What Do the Data Say? – Traps to Avoid

Examples that inform and educate

John Maindonald

11 April 2024

Contents

Preface

> It ain't what you don't know that gets you into trouble. Its what you know for sure that just ain't so. [Variously attributed; author unknown]

This booklet has as its intended audience, as well as practising researchers, anyone interested in using data as a basis for judgment. The critical processes and skills that are discussed have wide application, in everyday as well as in professional life.

The questions that data and data analysis may be asked to answer can often be stated simply. This may encourage the layperson to believe that the steps needed to provide answers are similarly simple. Very often, they are not. It is alluringly easy to create forms of data summary that misrepresent what the data have to say.
Or inadequate attention may be paid to the context from which the data have been taken. The context has to be understood and to feed into the way that the data are used, if conclusions are to be drawn that warrant credence.

The pages that follow avoid detailed discussion of methodology, instead focusing on what Kahneman, in his book "Thinking Fast and Slow"[1] calls "educating gossip", with examples taken from the media and from the research literature. Examples are chosen for the insight that they may provide, both those that show effective critical processes at work, and those that do not.

When they function well, scientific processes work to avoid the traps that will be discussed. They fail often enough that examples of failure are relatively easy to find. Cases where scientific processes have clearly failed are commonest where scientists work as individuals or in small groups with limited outside checks.

[1] Kahneman (2013)

In areas where the nature of the work requires cooperation between scientists with a wide range of skills, and where data and code are shared, those involved in the research provide a check on each other. Papers may be sent for comment to other researchers, or posted on the net for comment, prior to formal submission for publication. This allows informed and incisive criticism, with the more formal refereeing process providing supplementary checks.

Concerns about reproducibility, especially in wet laboratory biology and in psychology, have in the past decade attracted wide attention in the pages of Nature, Science, the Economist, psychology journals, and elsewhere. Among other needed changes, publication of experimental results needs to become a staged process that moves from "this is worth a look" to "has been independently replicated" to "established result".

Human psychology helps explain why humans so readily fall for conclusions that are both simplistic and wrong. Humans are programmed, by inheritance and by conditioning, to respond quickly to signs of immediate danger. Too often, we respond quickly and without careful thought in situations that call for a carefully reasoned response. Or, we may not have the skills needed for a carefully reasoned judgment. It is helpful to think about the psychology involved when humans make judgments as of two types, which Kahneman (2013) calls System 1 (jumping quickly to a conclusion) and System 2 (pondering). These are a useful starting point for drawing attention to common traps to which humans are prone. Those who know and understand common traps are better placed to avoid them.

Summary of points that are discussed

1. Understand the psychology — why the showcase of fallacies
   - Recognize and avoid the mental traps to which humans are prone.
2. Graphs can reveal surprises. They can also be drawn so that they deceive.
   - Choice of one or both scales such that equal relative changes mark out equal distances on the scale may show patterns that are not otherwise obvious.
3. From hindsight to insight — it is easy to be deceived.
   - If a sportsperson is at the top of their form, the only way to go is down. If at the bottom, the only way is up.
   - This is true also for success in business.
   - Chance, as well as form, obviously plays a part.
4. Medicine and health provides a good context into which to study a number of important issues. What causes cholera – bad air or bad water? Does drinking coffee help or harm health?
   - Randomized trials, if done rigorously, and participants closely reflect the population to which results will be applied, are a gold standard.
   - Think carefully about the outcome measure. Thus, is it "cancers" found? Or is it "deaths" within a stated timeframe.
5. Careful checks are needed when the attempt is made to use observational data to establish causation.
   - In group comparisons, adjustments are needed to account for prior differences between the two (or more) groups. What checks can be made that adjustments are adequate? It is in general impossible to be sure that adjustments account for all effects other than the effectm that is of interest.
   - In what direction, and through what causal chains, do causal effects go? Are people healthier because they exercise more? Or, do they

exercise more because they are healthier? Or, do causal effects go in both directions? What are the influences from dietary and other lifestyle factors?
- An overwhelming case can sometimes be made, as in the link between smoking and lung cancer, by bringing together multiple independent lines of evidence.

6. Mistakes that arise from statistical misunderstandings can have serious consequences.
   - This is the case also for fraudulent manipulation of data and/or evidence.
7. Weighting paradoxes appears in many different guises. A notable example is the Yule-Simpson paradox, which arises because we have naively added numbers in ways that give more weight to some combinations of effects than to others.
8. What direction does the correlation go?
   - Tall fathers are likely to have tall sons, but shorter than themselves.
   - What can be hard to grasp is that the effect goes on both directions. Tall sons are likely to have tall fathers, but shorter than themselves. The "regression to the mean" phenomenon goes in both directions.

   - Regression models require careful checking to ensure that they correctly reflect all systematic effects in the data.
   - The coefficient that is attached to an explanatory variable accounts for the effect of that variable when other variables are help constant. The coefficient is likely to change with the change to different explanatory variables that give the same predicted values.
9. Results become part of established science when they have survived informed critique. Work that claims to revise or overturn established results has itself to survive the same informed critique.
   - Uninformed critique is common. Watch for the tricks that detractors who find established scientific results inconvenient use in the attempt to discredit them.
   - When presented with scientific (or any) claims, ask/check whether claims have been carefully critiqued. For laboratory studies, have results been replicated?
   - Publication processes work best where authors contribute input and critique widely from across different disciplinary perspectives, and where data and code are made available for checking.

Chapter 1

Systems of human judgment

An understanding of the mental traps to which we are prone can help us avoid them. Kahneman's book (2013) *Thinking Fast and Slow* is a good starting point for thinking about the strengths and limitations of human thinking processes. There is no good substitute for the use of "educating gossip", as Kahneman describes it, for training in effective judgment and in decision making.

Important themes that Kahneman notes are

- We have "an excessive confidence in what we think we know".
- We too readily judge decisions by outcome, rather than by the strength of the arguments that support them.
- We have two selves — an experiencing self and a remembering self
  - These do not always have the same interests.
- Automatic memory formation has its own rules
  - This is exploited to improve the memory of a bad episode
- "We easily think associatively, … metaphorically, casually, but statistics requires thinking about many things at once …"

1.1 System 1 and System 2 — further comments

Humans have been conditioned to respond quickly to immediate risks and challenges, without stopping to consider too carefully whether what we heard was a false alarm. They also have the ability, when the occasion seems to demand it, to stop to ponder. This is the basis for Kahneman's categorization of human

thought processes as of two types — System 1 which jumps rapidly to make a judgment, and System 2 which takes time for careful consideration.[1]

System 1 Features are

- It may answer an easier question in place of a harder.
- It responds to irrelevancies — priming, framing, affect, memory illusions, illusions of truth, …
  - Priming by one stimulus can affect the response to a second stimulus that occurs shortly aferwards
  - The portrayal of logically equivalent alternatives in different ways or 'frames' can affect the response
  - The emotion or 'affect' generated by a question can affect the response
- It has little understanding of logic and statistics
- It cannot be turned off, but it can be trained
- When flummoxed, it calls on System 2

System 2 features are

- It keeps you polite when angry, alert when driving in a severe rainstorm
- In its world, gorillas do not cross basket-ball courts …[2]
- Problems that put 1 & 2 in conflict may require large mental effort & self-control to overcome the impulses and intuitions of System 1.
- Its effectiveness depends, in important areas, on training.

Both systems are amenable to training. A well-trained System 2 helps greatly in creating a better System 1. Further points are:

- Healthy living is a compromise
  - Recognize situations where mistakes are likely
  - Aim to avoid significant mistakes when the stakes are high
- Untrained humans are poor intuitive statisticians
- Judgments about statistical issues may require us to think about more than one, even many, things at once.
- Too often, we jump to conclusions, without careful assessment.
- We may not be equipped to make an informed and carefully thought through decision.

[1] See https://suebehaviouraldesign.com/kahneman-fast-slow-thinking/ for further comment.

[2] See http://theinvisiblegorilla.com/gorilla_experiment.html

1.2 The Intuition of Professionals

Effective professional training is designed to ensure that at least some of the results of well-tuned System 2 expert judgment operate at a System 1 level. The professional will, if the training is doing its job, build up a repetoire of System 2 judgments that will later, when the circumstances seem to demand it, be available at a System 1 level.

> The situation has provided a cue; this cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition. (Simon 1992, "What is an Explanation of Behavior?")

Obstacles to effective judgment

Even those who are experts in their field can be similarly prone to judgments that have no foundation in fact. The following comment appeared in a discussion of the response to a U.S. Preventive Services assessment that prostate screening, when used in accordance with then current treatment practices, was doing more harm than good.[3]

> Even faced with … evidence … from a ten-year study of around 250,000 men that showed the test didn't save lives, many activists and medical professionals are clamoring for men to continue receiving their annual PSA test.

New evidence emerges as time proceeds, and there are advances in the approach to treatment. At least part of the problem has been a rush to treatments that themselves risk increasing damage and the risk of death. Note the comment in Brawley (2018) that

> Over the past few years, the benefit-to-harm ratio has improved in favor of benefit if the man understands that active surveillance may be a reasonable path if diagnosed.

1.3 A demand for discipline & careful thought

- We make judgments based on evidence that is too limited
- We are easily fooled by irrelevancies

[3] Association of Professional Psychologists, web post on Arkes and Gaissmaier (2012)

  - Kahneman has brought together evidence on what & how.
- Even when data are there for the taking, someone has to notice, to collate the data, and to understand its uses and limitations
- Randomized controlled experiments (RCTs) are often the ideal, but require meticulous planning. If effects of interest are small, the numbers required may be very large. See further, Subsection 4.2
  - A limitation is that results apply only to the population from which trial participants were taken. Any wider generalization has to be justified
- Observational data does not easily, if at all, substitute for the use of RCTs. It is in general impossible to be sure that all sources of bias have been properly accounted for
- Keep in mind Yule-Simpson "paradox", which we will encounter later in Subsection 6
  - The paradox lies in the failure of human intuition to accommodate straightforward arithmetic!

1.4 Further examples

The "conjunction fallacy"

This has also, as a result of the example given in Tversky and Kahneman (1983) come to be known as "the Linda problem". The name 'Linda' comes from the question and usual response that are given by way of example.

> Linda is a 31-year old philosophy graduate, single, outspoken, and bright. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which of the following is more probable?

- Linda is a bank teller.
- Linda is a bank teller, and active in the feminist movement.

Adding the further descriptor "active in the feminist movement" can only lower the probability, or just possibly leave it unchanged. Instead of assessing the balance of probabilities, we are tempted to ask which description best meshes with what we have been already told about Linda.

"Linda is active in the feminist movement" is the single descriptor" that respondents see as best fitting Linda. While that was not what was asked, one has to pay close attention to prevent System 1 from substituting that for the question

that was asked. Note that the correct answer will be "a bank teller", irrespective of the way that Linda was characterized before the question was asked.

In part, the issue is one of use of language. The "correct" answer is asking us to use the word "probable" in a strict technical sense.

Even careful critics sometimes get it wrong

An irony is that Kahneman was, as he has acknowledged, himself fooled into taking at face values papers that claimed to show that verbal concepts could have the effect of altering behaviour. Thus

- Being asked to write down stories about unethical deeds made people more likely to want to buy soap;
- Subtly drawing attention to money, e.g., leave banknotes lying around, made people feel more self-sufficient, and care less about others;
- Priming people with old age related words leads people to walk more slowly away from the lab as research assistants armed with stopwatches timed their movements.

As Ritchie (2020) notes (p.28) Kahneman was not alone in being fooled — the study about priming with old age related words has been extensively cited in psychology textbooks. None of these claims have stood up in attempts at replication, with larger numbers and with greater care to avoid unconscious sources of bias. Thus, in the replication of the study relating to age-related words, infrared beams were used to measure time taken to walk between two points in a hallway, rather than research assistants who knew the group to which participants had been assigned.

Think again — a very simple example

Is symptom X associated with disease A?

	Has Disease	No Disease
Symptom X Present	20	10
Symptom X Absent	80	40

The symptom occurs with the same relative frequency, whether or not a person has the disease. Nisbett comments that most people, including nurses and doctors, interpret such evidence wrongly (Nisbett 2016, 129–30).

A test to check understanding of risk

See the 2-minute testm "Do you understand risk?".[4]

1.5 Misbehaving humans!

The discipline of "behavioural economics" largely took shape as a result of the work of Richard Thaler. Kahneman was one of two mentors who strongly influenced Thaler — the other was Amos Tversky, who had worked closely with Kahneman.

Thaler and Ganser (2015) explores the extent to which humans do not behave like the rational agents of classical economics, agents to whom Thaler gives the name "econs". Added to the irrationality with which we often act is that our personal priorities are unlikely to align precisely with those of econs.

Note also comments in Part 4 of Kahneman (2013), titled "A conversation with the discipline of economics". In a discussion on "The Prospect theory model of choice", Kahneman comments on

- "… unfortunate tendency to treat problems in isolation"
- Framing effects — inconsequential features shape choices

Hence, "a challenge to the assumptions of standard economics".

1.6 Negotiating Life in an Uncertain World[5]

Questions that it can be helpful to ask include

1. Risk of what? (Showing a symptom, …, Death)
2. What is the time frame? (next 10 years, or lifetime)
3. How big is the risk? (Look at risk in absolute terms)
4. Does the risk apply to me? (Age, sex, health, …)
5. What are the harms of "finding out"? (False alarms, invasive diagnostic procedures, unnecessary or dangerous treatments.)

[4] http://www.riskliteracy.org/

[5] These questions came originally from the Harding Center web site https://www.hardingcenter.de/en.

Chapter 2

Effective use of graphs

2.1 General principles

- Focus the eye on features that are important
- Avoid distracting features
- Lines that are intended to attract attention can be thickened
- Where points should be the focus, make them large & dark
  - It often makes sense to de-emphasize the axes.
- If points are numerous and there is substantial overlap, use open symbols, and/or use symbols that have some degree of transparency.
- Different choices of color palettes are effective for different purposes.
- Bear in mind that the eye has difficulty in focusing simultaneously on widely separated colors that are close together on the same graph.

2.2 Varying time intervals — show rates, not counts

A graph that was essentially the solid segmented solid line in Figure 2.1 appeared in National Science Foundation (1975) "Science Indicators, 1974". The segmented line gives a highly misleading impression for the four years 1971-1974, as opposed to earlier points, where numbers are totals over decades. It joins up a final point that is a different measure from earlier points.

The gray dots, and the axis on the right, show rates per year, thus comparing like with like.

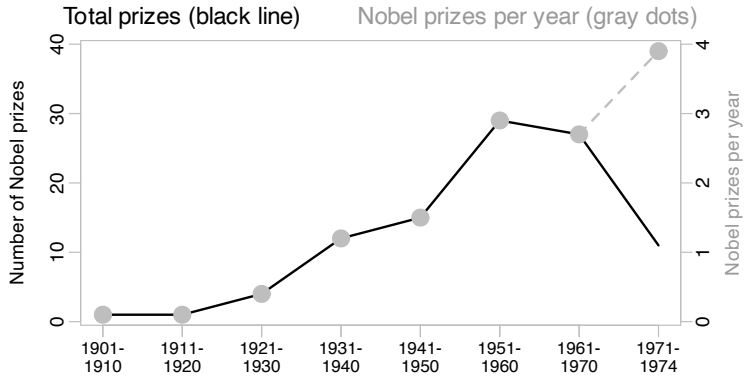

Figure 2.1: The black line shows numbers of US Nobel prizes, for given time intervals. The gray dots. with the right axis label, show average per year.

The same principle applies for intervals of measures other than time — for example of length or volume.

2.3 Banking — the importance of aspect ratio

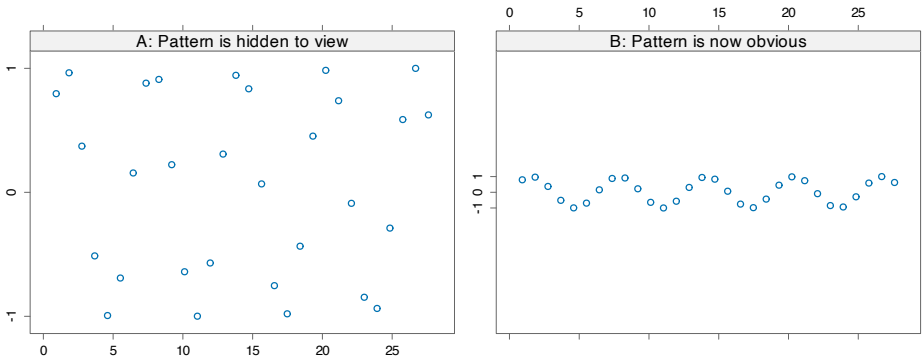

Figure 2.2: The same data are used for both graphs. The pattern that is not obvious in Panel A is very obvious in Panel B

Patterns of change on the horizontal scale that it is important to identify should bank at an angle of roughly 45° above or below the horizontal. Angles in the approximate range 20° to 70° may be satisfactory, and the aspect ratio should be chosen accordingly.

2.4 Scales that show changes by equal multipliers

Figure 2.3 shows two plots of the same data. Panel A plots brain weight (grams) against body weight (kilograms), for 28 "animals". Panel B plots the same data, but now equal distances on each scale show changes by the same factor (i.e., change in relative weight).

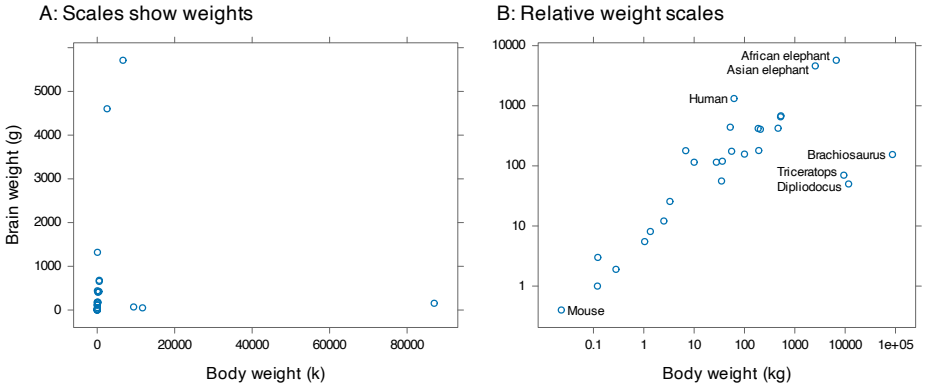

Figure 2.3: Panel A plots brain weight (grams) against body weight (kilograms), for 28 'animals'. Panel B plots the same data, with relative weight scales, i.e., equal distances on each scale show changes by the same multiplier.

Often, when measurement data span a large range (e.g., a change from smallest to largest by a factor of 100 to 1 or more), it is a relative amount scale that is appropriate.[1]

[1]Technically, such scales are termed logarithmic, as opposed to straight line or linear. A logarithmic transformation is used to obtain such relative distance scales.

2.5 Different graphs serve different purposes

The line in Figure 2.4A shows the broad overall pattern, while Figure 2.4B shows how that pattern needs to be tweaked to more closely reflect the data.

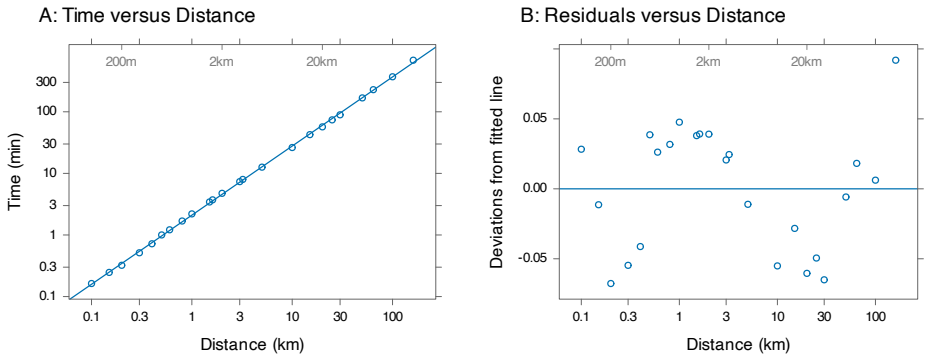

Figure 2.4: Panel A plots world record Time (as of 2006, in minutes) against Distance (in kilometers), for field races. On both the 'x' and 'y' axes, the scale is one on which equal distances on the scale correspond to equal relative changes. Panel B plots deviations from the fitted line in the 'y' direction, otherwise known as residuals, against Distance. The deviations are approximate relative differences from the line. Thus a 0.05 difference is a difference that amounts to 5% of the time predicted by the line.

Notice, in Panel A, the use of scales for which which equal distances on the scale correspond to equal relative changes. This is achieved by specifying logarithmic scales, on both axes. There is a loglinear, i.e., straight line on logarithmic scales, relationship.

In Figure 2.4, the line looks to be a good fit. The range of times is however large, from just under 10 seconds to close to 11.5 hours. All except the largest difference from the fitted line are a less than 7% change, and are not at all obvious in Panel A. There is a very clear pattern of systematic differences in Panel B that reflects differences in human physiology, very likely between the athletes who excel at the different distances.

The line can be interpreting as implying a 13% increase in the time per unit distance for every unit increase in the distance. The units may for example be

units of 100 meters, or kilometers. Panel B indicates that the pattern of increase moves down to a local minimum at around 200 meters, up to a local maximum at around 1 kilometer, down again to a local minimum at around 20 kilometers, and then steadily up again.

Relative distance scales

Figure 2.5 shows different "equal physical distance along the scale" labels that might be used for the relative `Distance` ("logarithmic") scale in Figure 2.4 in Subsection 2.5.

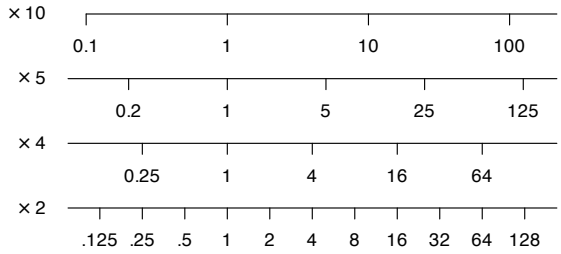

Figure 2.5: Different labelings, all with tick marks at the same relative distance apart, are shown for the 'Distance' scale. The multipliers for the 'Distance' values that are plotted are, starting at the bottom, 2, 4, 5, and 10.

2.6 Helpful web links are:

- Good & bad graphs (Ihaka, lecture notes)[2]
- Misleading graphs[3]
- Color Brewer[4]

[2] https://www.stat.auckland.ac.nz//~ihaka/120/Lectures/lecture03.pdf
[3] https://www.statisticshowto.com/misleading-graphs/
[4] https://colorbrewer2.org/

Chapter 3

Selection and survivor bias

In mind here are cases where the data are not a random sample.

3.1 The hazards of convenience samples

Quota sampling has often been used as an alternative to random sampling — quotas are set for age categories, male/female, and socioeconomic categories that are designed to ensure that the sample is representative of the wider population. In polls prior to the 1948 US presidential election that pitted democrat Harry Truman against republican Thomas Dewey, pollsters were given strict quotas, but otherwise left free to decide who they would approach. Polls by three different organizations gave Dewey a lead of between 5% and 15%. In the event, Truman led by 5%.

Convenience samples sometimes have a story to tell

This is not to rule out all use of convenience samples. Convenience samples, taken within a limited population, can sometimes be useful in setting a bound. It is strongly in the public interest that scientists have reasonable freedom for responsible expression of their minds on issues of public concern. In an informal 2015 survey, 151 Crown Research Institute scientists (out of 384 who responded) answered yes to the question "Have you ever been prevented from making a public comment on a controversial issue by your management's policy, or by fear of losing research funding?" The 384 who responded will undoubtedly be a biased sample. Irrespective of the size of the bias, the number who had not

been allowed to speak their mind was large enough to be a cause for serious concern. Hon Joyce's response, to the effect that as this was not a scientific survey of all CRI scientists (to this extent, true), its evidence of large concern could be ignored, was an evasion. Equally disturbing was the reaction of the NIWA management, suggesting that they did not accept a responsibility to defend transparency.[1]

3.2 UK cotton worker wages in the 1880s

Prior to the Boot and Maindonald (2008) paper[2] the main source of published information on cotton worker wages in the UK in the late 19th century were results from an 1889 US Bureau of Labor survey, intended for use for comparison with the US cotton industry wages.

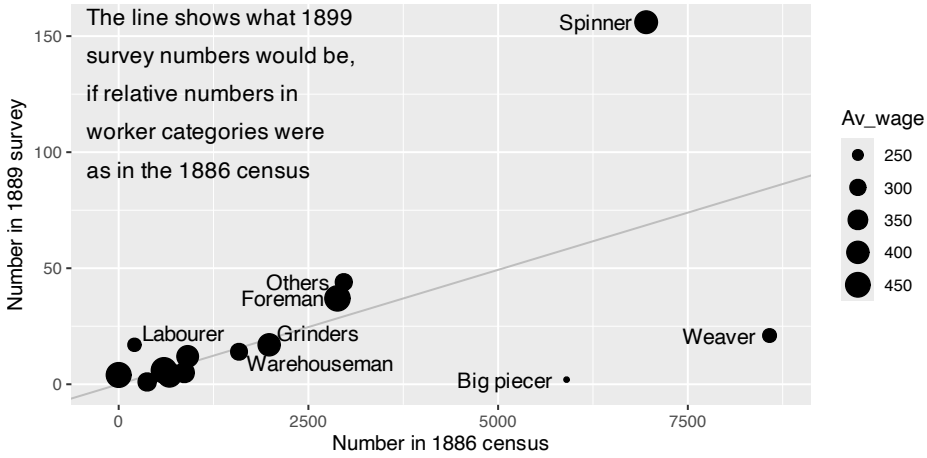

Figure 3.1: Cotton worker wages in the UK — 1889 US Bureau of Labor 'survey' versus 1886 census data. Wages are in pence per week.

Figure 3.1 compares the US Bureau of Labor survey numbers with 1886 census numbers of different types of full time UK cotton operatives. The 1889 survey

[1] See https://sciblogs.co.nz/infectious-thoughts/2015/08/28/niwa-in-astonishing-attack-on-scientist-association/

[2] "New estimates of age- and sex- specific earnings and the male-female earnings gap in the British cotton industry, 1833-1906"

shows some strong biases — a result it would seem of geographical bias and of the informal data collection methods that were used. The high wages given to spinners were grossly over-weighted in the US Bureau of Labor survey, while Big Piecers and Weavers were grossly under-represented. A guess is that workers were asked for information on their wages as they left work, and that the survey personnel happened to catch employees at a time when there was a large preponderance of highly paid spinners, and an untypically small number of big piecers and weavers. The net effect was a gross over-estimate of average wages.

3.3 The uneasy path from hindsight to insight

There is a mix of selection and survivor bias when data from the past are used as a guide to the future, with no allowance for the source/target difference. The target about which we wish to make judgments lies in the future, while the data are from the past. Think about a business that is planning for the future. One can never know, until after the event, all the ramifications of the choices made.

Businesses may be selected as examples of effective business practice because they were, at the time when the data were collated, successful. Likewise, it is athletes who have been successful in the recent past who are likely to be selected to appear on the covers of sports magazines. In either case, this gives a biased picture of what can be expected in the future — in many cases those who are picked out will be close to the peak of their success, and/or have had unusual luck. High levels of success in the recent past will not always translate into success in the following year or years. How often will past success translate into future success? In order to discover, it is necessary to collect relevant data.

Tales of standout past success

Collins (2001), in the book *Good to Great*, identified 11 companies, from 1,435 studied, as standouts. Since 2001, 5 have performed better than average, and 6 worse. Notable later changes in fortune were

- Fannie Mae — 2001: >$80 per share: 2008: <$1
- Circuit City: Bankrupt in 2009

Waterman and Peters (1982), in the book *Close to consumer"*, identified 43 successful companies, as having a "bias for action", and being "close to consumer". From the 35 that were publicly listed, Smith (2014) noted that 15 had done better than average, and 20 worse. See Smith (2014) for further commentary.

In all cases, companies that were chosen as examples of standout success were likely to be near to the peak of their performance, as judged by Collins, or by Waterman and Peters. Overlaid on this is the regression effect that will be discussed in Chapter 8.

3.4 The message in the missing bullet holes

In World War II, the US air force was concerned that too many of their planes were being shot down. What were the priority areas to which protective armour should be added, given that the extra weight meant that they could not be placed everywhere? Abraham Wald's insight was that survivor bias was to be expected, with the density of bullet holes providing evidence about the extent of bias, and the implications for identifying the part(s) of the planes that would benefit most from additional protection. See Abraham Wald and the Missing Bullet Holes[3], which is an excerpt from Ellenberg (2015).

The numbers of hits per square foot were:

Engines	Fusilage	Fuel system	Rest of plane
1.11	1.73	1.55	1.80

Wald argued that the gunshots were likely to have been spread very nearly uniformly over the planes as a whole, those that were shot down, and those that survived, The reason for relatively fewer bullet holes in the engine and fuel system areas was that hits in those areas were more likely to bring the plane down, so that they did not return.

[3] https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d

Chapter 4

Medicine and health

A focus on medicine and public health is used as a context in which to introduce ideas and issues that are more generally relevant. There is wide acceptance that the evidence provided by randomized controlled trials (RCTs) that are conducted to high standards is at the top of a hierarchy of evidence.

Web sources are noted that, because they base their advice on careful and transparent evaluations of the available evidence, can be trusted.

4.1 Useful sources of advice and information

Here are noted web resources that, for many issues of major interest, provide carefully collated assessments that are based on a critical examination of the whole range of evidence that the authors could identify. There may be many published studies, not all of equal quality, that bear on a medical issue. Or there may be one, or a small number, of large-scale high quality trials that are used as a basis for judgment. It is important that assessments are updated as new evidence emerges.

Resources that will be noted are

- Harding Center for Risk Literacy
- The Cochrane Center
- Winton Centre for Risk and Evidence Communication

Note also the US Consumer health information site

The Harding Center for Risk Literacy

There is extensive informative content on the Harding Center for Risk Literacy site.[1] Note in particular

- Medical Fact Boxes. Look under **Transfer and Impact | Fact Boxes**
- Covid-19. Look under **Transfer and Impact | Corona pandemic**
- Consumer Empowerment. **Transfer and Impact | Consumer Empowerment**
- Risk and Evidence Communication. Look under **Research | Risk and Evidence Communication**
- VisRisk - Visualization and Communication of Complex Evidence in Risk Assessment. Look under **Research | VisRisk**
- Horizon2020 Project FORECEE (on balancing female cancer risk against the risk arising from false-positive alarms and overdiagnosis). Look under **Research | FORECEE Project**
- Drone Risks Project. Look under **Research | Drone Risks Project**
- Publications. Look under **The Harding Center | Publications**
- In the Media. Look under **The Harding Center | In the Media**

Harding Center Medical Fact Boxes

Medical fact boxes[2] provide visual and tabular summaries of the current "best" evidence, from randomized controlled trials. The comparison may be with a placebo, or with an alternative that is known to be effective. Detailed references are given. Where available, reliance is on Cochrane studies. Available fact boxes, as of March 2024, appear under the headings:

- Vaccines
- Low back pain
- Antibiotics
- Early detection of cancer: breast, colon, prostate, ovarian
- Cardiovascular Diseases
- Osteoarthritis of the knee
- Tonsil surgery
- Pregnancy and childbirth

Scroll down the home page[3] to find the headings **Risk Literate** (with a link

[1] https://www.hardingcenter.de/en
[2] https://www.hardingcenter.de/en/transfer-and-impact/fact-boxes
[3] https://www.hardingcenter.de/en

to a risk quiz), and **Quick Risk Test** (with a link to a test that is targeted at medical students and medical professionals. On the development of this latter test, see Cokely et al. (2012).

The Cochrane Center[4]

The Cochrane Center's mission is "to promote evidence-informed health decision-making by producing high-quality, relevant, accessible systematic reviews and other synthesized research evidence." They rely heavily on meta-analyses, looking for the balance of evidence across all relevant studies.

4.2 Randomized Controlled Trials vs other study types

Two types of study are widely used in medical and other contexts — randomized controlled trials (RCTs), and population-based studies. These can, in both cases, be broken down into further sub-types. There may be elements of both these types of studies.

Randomized Controlled Trials — the gold standard?

For simplicity, it will be assumed that there are two alternatives to be compared, in what is known as a two-armed trial. In medical trials the two arms may be a treatment and a placebo, with the placebo (something harmless that has no effect) made to look as similar as possible to the treatment.

An important distinction is between pre-clinical trials, often with animals such as mice, and clinical trials with human subjects. Pre-clinical trials are likely to be conducted with a small number of mice or other animals, and are intended to check whether a drug or other treatment warrants further investigation. Evidence will be presented in Section 9.4 that suggests that these trials are commonly not achieving their intended purpose.

Clinical trials typically are conducted in three phases. Phase I applies the treatment to a small number of subjects and checks that the treatment appears safe. Phase II, perhaps with several hundred subjects, looks for evidence of an effect, and how this might relate to dose level. Phase III trials are conducted with large numbers of subjects, typically spread over multiple treatment centres, and are designed to check whether the treatment is effective, what side effects there may be, and whether there are issues with particular subgroups.

[4] https://www.cochrane.org/

Important issues are

- Use a random mechanism to assign to assign to treatment as against control — in a medical screening study to "screen" or "not screen"
  - The aim is to ensure that apples are compared with apples
  - Treatment and control must otherwise be treated in the same way.
- There must be strict adherence to a protocol
  - Minor departures that may, e.g., allow unconscious bias in the way that results from the different groups of participants are measured, can invalidate results.
  - In clinical trials an ideal, not always possible, is the double blind trial where neither the individual nor the clinician involved knows which drug (or other treatment) the individual has received.
- Results apply, strictly, to those who meet the trial entry criteria
  - This may limit relevance to the general population

Especially in medical trials, think carefully about the outcome measure. It is not enough to show that a screening program will pick up otherwise undiagnosed cancers.

- In a screening trial, e.g., for prostate cancer, there are risks both for those who test positive, and for those who test negative.
  - The process used to check for cancer may itself bring a smaller or larger element of risk.
  - Positives may be false positives, leading to more invasive checks which may themselves carry a risk. Thus, for prostate cancer, a positive PSA test is likely to lead to a biopsy that itself has been estimated to carry a 5% risk of serious side effects, with a much higher proportion of less serious effects (Levitin 2015, 245)
  - Some slow growing cancers may be better left untreated, rather than exposing the patient to a treatment that may itself do serious damage.

For a helpful animated summary of some of the key issues, see: https://www.youtube.com/watch?v=Wy7qpJeozec. Pashayan et al. (2020) provide an overview of progress towards the personalized early detection and prevention of breast cancer, noting priority areas for action.

A note in passing: HiPPO decisions vs A/B testing

Randomized studies are widely used outside of medicine. Randomization is a key component of the way that Google and others test out, e.g., the effect of

different web page layouts.

- HiPPO = "Highest paid person in the Office."
- The term "A/B testing" is sometimes used to refer to randomized testing of alternatives.

A/B testing helped propel Obama into office! An experiment was conducted that involved 15 million people, or about 25%, from its email list. The signup forms had one of nine different combinations of images with words on which recipients were invited to click, thus:

	Learn more	Join us	Sign up now
Obama photo	✘	✘	✘
BW photo of Obama family	✔	✘	✘
Obama speaking	✘	✘	✘

The black and white photo of the Obama family, with the words "Learn more", generated the most clicks.

Young (2014) gives an account of A/B testing as it might be used for improving library user experience.

Population studies — groups must be broadly comparable

- Adjust prunes to look like apples (is it possible?)
- Can one ever be sure that the adjustments do the job?
- Potential for biases is greater than for randomized controlled trials.

Where a treatment is compared with a control group, the idea is to use a regression type approach to adjust for differences in such variables or factors as age, sex, socioeconomic status, and co-morbidities. "Propensity score" approaches try to summarize such group differences in a single variable (or, in principle, two or more variables) that measure the propensity to belong to the treatment as opposed to the control group. While their effectiveness for this purpose may be doubted, they can be used to provide insightful graphs that check the extent to which the groups are broadly comparable on the variables and/or factors used to adjust for differences.

The generally negative view of observational studies that is presented in Soni et al. (2019) (studies in oncology) contrasts with the more positive view offered in Anglemyer, Horvath, and Bero (2014) (health care), for observational studies that have been conducted with high methodological rigour. The strongest evi-

dence comes, as with the link between smoking and lung cancer, from multiple studies, with different likely biases, that all point strongly in the same direction.

Issues for all types of study

What are the relevant outcome measures?

- e.g., cancer – malignancies found & removed, or deaths
  - deaths from cancer, or from all causes (for some individuals, the treatment may be more damaging in it medium to long term effect than the cancer)

Care is required to deal with survivor, as well as other, biases.

4.2.1 False Positives

In contexts where the number of false positives is likely to be high relative to the number of true positives, screening programs may have serious downsides that outweigh the benefits.

Excess iron syndrome, known as "haemochromatosis", affects around 1 in 200 in the New Zealand population. Consider a test that has an 80% accuracy, both for detecting the syndrome among those who have it, and for not detecting among those who do not.[5]

Excess iron syndrome, known as "haemochromatosis", affects around 1 in 200 in the New Zealand population. Consider a test that has an 80% accuracy, both for detecting the syndrome among those who have it, and for not detecting among those who do not.[6]
Among 2000 tested (10 with and 1990 without)

- there will on average be **8** out of 10 true positives
- the 1990 without the syndrome will split up as detected +ve to detected -ve in an 20%:80% ratio. Thus there will be, on average, $.2 \times 1990 = \mathbf{398}$ false positives.

Overall, those detected as positive split up in a true:false ratio of 8:398, i.e., 8/406 or just under 2% of the positives are false positives. If all positives were detected as positive, the 8/406 would change to 10/406.

[5] These are known as the "sensitivity", and the "specificity".

[6] These are known as the "sensitivity" (true positive rate"), and the "specificity" (true negative rate).

A test with this kind of accuracy becomes much more useful in a subset of the population already known to be at high risk, perhaps as identified by a genetic test, or perhaps because of a medical condition commonly associated with the syndrome.

4.3 Hierarchies of evidence

There is broad agreement among medical researchers on the hierarchy of evidence that is set down in the (US Preventive Services Task Force 1989) guide:

- Level I: Evidence obtained from at least one properly designed randomized controlled trial.
- Level II-1: Evidence obtained from well-designed controlled trials without randomization.
- Level II-2: Evidence obtained from well-designed cohort studies or case-control studies, preferably from more than one centre or research group.
- Level II-3: Evidence obtained from multiple time series designs with or without the intervention. Dramatic results in uncontrolled trials might also be regarded as this type of evidence.
- Level III: Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

The CONSORT 2010 statement (Schulz, Altman, and Moher 2010) sets out detailed criteria for assessing randomized controlled trials (RCTs). For Level II studies, the STROBE guidelines (Erik von Elm et al. 2007) set out reporting standards.

Use of such criteria is essential when evidence that is available from multiple randomized controlled trials, perhaps supplemented by evidence from studies at lower levels of the hierarchy, is brought together in a systematic review. Evidence at level II, and especially at level II-3, should ideally be checked by conducting an RCT. This is not always possible, for ethical or practical reasons.

Evidence from one type of study may complement evidence from another. A paper entitled “Resolved: there is sufficient scientific evidence that decreasing sugar-sweetened beverage consumption will reduce the prevalence of obesity and obesity-related diseases” (Hu 2013) provides an example. Hu brings evidence from short-term randomized controlled trials together with evidence from long term cohort studies (4 or 8 years) to make a convincing case.

Clinical trials have their own problems and issues. Using evidence from pub-

lished review sources, Chalmers and Glasziou (2009) found issues with the choice of research questions; the quality of research design and methods, and the adequacy of publication practices. They reported that 50% of studies were designed without reference to systematic reviews of existing evidence, and that 50% were never published in full.

Planning and execution failures are set in stone by the time that a research report is sent for review. Pre-registration, involving the depositing a research question and study design with a registration service or journal before starting an investigation, allows peer review feedback that can elicit suggestions for improvement and detect any potential flaws before the study begins.[7] Even more than for industrial quality control, processes are needed that prevent defects from appearing in the first place, with screening for defects at the end of the production line used as a check that those processes have done the job asked of them.

Chapter 5 will discuss issues for using observational data as a basis for inferences.

4.4 Avoid, or expose infants to peanuts?

Clinical practice guidelines introduced in or around the year 2000 had "recommended the exclusion of allergenic foods from the diets of infants at high risk for allergy, and from the diets of their mothers during pregnancy and lactation."

It was then a surprise to find that the prevalence of peanut allergy has substantially increased in the recent past, doubling in Europe between 2005 and 2015, suggesting that advice given to parents of young children to avoid foods containing peanuts may have been counterproductive. This reassessment was supported, at least for infants who at four months had either severe eczema or food allergy or both, and thus were at high risk of developing a peanut allergy, by the LEAPS study reported in Du Toit et al. (2015).

As noted, the LEAPS study was limited to infants who at four months had either severe eczema or food allergy or both. Infants were stratified into two groups following a skin-prick test, with each group then randomized between those exposed to peanut extract, and those not exposed.

Among 530 infants in the population who initially had negative results on the skin-prick test, the prevalence of peanut allergy at 60 months of age was 13.7%

[7] See https://plos.org/open-science/preregistration/

(37/270) in the avoidance group and 1.8% (5/272) in the consumption group.[8] Among the 98 participants who initially had positive test results, the prevalence of peanut allergy was 35.3% (18/51) in the avoidance group and 10.6% (5/47) in the consumption group. There was no between-group difference of consequence in the incidence of serious adverse events.

In both groups, numbers and percentages are for those who were assigned to the group and whose results could be evaluated, whether or not they followed the treatment protocol to which they were assigned. In technical terms, these are results from an "intention to treat" analysis. Such an analysis is designed to mirror what can be expected in practice — not everyone who starts off in one group will stick to it. It answers questions about what to do with subjects who did not fully follow the treatment to which they were assigned.

The results were followed, in 2016, by changes to guidelines that recommended introduction of peanut and other allergenic foods before 12 months. The assumption that avoiding early exposure to peanuts would reduce risk of later development of peanut allergy was, it was judged, likely wrong for all infants.

4.5 The effectiveness of surgery – RCTs are challenging

The blurb on the back cover of Harris (2016) states that

> For many complaints and conditions, the benefits from surgery are lower, and the risks higher, than you or your surgeon think.

Humans are very prone to the *post hoc, ergo propter hoc* fallacy: "it followed, therefore it was because of" fallacy. Harris argues that unless the benefits of a surgical procedure are clear, the only ethical way forward is to do a randomized trial where the procedure is compared with a sham procedure. Such trials are not easy to design and execute. Nonetheless, there are a number of important cases where such comparisons have been made.

Bloodletting is a prime example of a surgical procedure that has faded away due to evidence, not just of a lack of effectiveness, but of serious harm.[9] The practice attracted widespread debate in the 19th century and into the early 20th century, with its defenders making such claims as

[8] There were twelve further infants in this group whose results could not be evaluated.
[9] For the history, see for example Seigworth (1980)

> "blood-letting is a remedy which, when judiciously employed, it is hardly possible to estimate too highly"

A comment in Watkins (2000) is apt

> Medical evidence is trusted, and we must retain that situation and ensure that it is not abused. It is possible to be an extremely good doctor without being numerate, and not every eminent clinician is best placed to give epidemiological evidence. Doctors should not use techniques before they have acquainted themselves with the principles underlying them.

What are the implications for medical practice?

4.6 Screening for cancer — how relevant is historical evidence

Screening for cancer is an area where, if the interest is in risk of death, it is necessary to wait for perhaps several decades before there is a high enough number of deaths that results can be usefully evaluated. Changes that can be expected in the interim include:

- Screening may become more sensitive, perhaps picking up a higher proportion of relatively benign cancers that are unlikely to ever have serious effects.
- There may be an increased ability to distinguish between benign and more aggressive cancers.
- More effective and/or less invasive treatments may become available, and earlier treatments finessed.

All cause death rates are a more relevant measure than cancer specific death rates, as treatments may themselves have harmful effects. Whether or not results from clinical trials in past decades remain relevant to current circumstances, their results do highlight important questions.

PSA Screening for Prostate Cancer, & more

Numbers (rounded) in the following table are from a Harding Center fact box. They are for men 50 years or older who either did or did not participate in

prostate cancer screening, using the PSA test, for 16 years.[10]

	1000 men, No screening	1000 men, Screening
Deaths (prostate cancer)	12	10
—	—	—
Biopsy & false alarm	0	155
Unnecessary treatment	0	51

About 10 out of every 1,000 men with screening, and 12 out of every 1,000 men without screening, died from prostate cancer within 16 years. This means that 2 out of every 1,000 people could be saved from death from prostate cancer by early detection using PSA testing. Deaths from any cause were around 322 in both groups.

Numbers for benefits are based on four studies with about 77,000 participants (progressive cancer), four studies with about 472,000 participants (overall mortality), and eleven studies with about 619,000 participants (prostate cancer specific mortality). The numbers for harms are based on seven studies with approximately 128,000 participants (false-positive results within three to six participations in PSA testing for early detection) and nine studies with approximately 274,000 participants (over-diagnosis and over-treatment). See the web site for references to the studies.

Unlike the biopsies that may follow a positive PSA test, the PSA test has no direct potential to cause physical harm. Harm results from an undue readiness to use the test result as a reason for further potentially harmful testing and treatment. “Wait and watch” is often the preferred strategy.

See Martin et al. (2018), Levitin (2015)

*ch.*6

, Fung (2020, 278–81), the web page How Patients Think, and How They Should[11], and regularly updated summary of the evidence can be found at PDQ Cancer Information Summaries.

[10] https://www.hardingcenter.de/en/early-detection-of-cancer/early-detection-of-prostate-cancer-with-psa-testing

[11] https://www.nytimes.com/2011/10/09/books/review/your-medical-mind-by-jerome-groopman-and-pamela-hartzband-book-review.html

Breast cancer screening

The Raichand et al. (2017) review starts with the comment:

> The recent controversy about using mammography to screen for breast cancer based on randomized controlled trials over 3 decades in Western countries has not only eclipsed the paradigm of evidence-based medicine, but also puts health decision-makers in countries where breast cancer screening is still being considered in a dilemma to adopt or abandon such a well-established screening modality.

The short summary, last updated in October 2019, from the Harding Center Fact Box for Mammography Screening, referring to women 50 years (a few trials looked at women aged $ $40) and older who either did or did not participate in mammography screening for approximately 11 years reads

> Mammography reduced the number of women who died from breast cancer by 1 out of every 1000 women. It had no effect on the number of women who died from any type of cancer. Among all women taking part in screening, some women with non-progressive cancer were over-diagnosed and received unnecessary treatment.[12]

The chief English source of evidence for the fact box is the Cochrane review Gøtzsche and Jørgensen (2013). The eight eligible trials included more than 600,000 women aged between 39 and 74, all reported between 1963 and 1991. One trial was excluded because the randomization had not produced comparable groups. Four trials had inadequate randomization. The three trials with adequate randomization did not find an effect of screening on total cancer mortality.

Løberg et al. (2015) provide a slightly more detailed breakdown of the evidence, as applied to women who were screened for 20 years, starting at age 50, with mortality assessed at ages 56 to 75 in the UK. Figure 4.1 is a visual summary. Interval cancers are cancers that are detected in between regular screens. A prior normal screen may give a false assurance and lead to a delay in seeking help when symptoms appear. See also the regularly updated summary of the evidence atPDQ Cancer Information Summaries.

[12] https://www.hardingcenter.de/en/early-detection-breast-cancer-mammography-screening

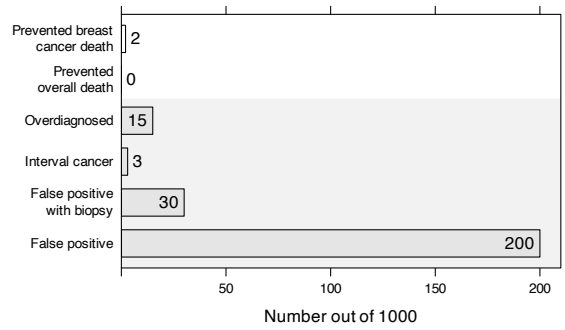

Figure 4.1: Estimates of benefits and harms of screening, as applied to the observed incidence of invasive breast cancer (women aged 50 to 69 years) and mortality (women aged 55 to 74 years) in the UK in 2007.

The area by area phasing of the introduction of a mammography screening program in Ireland over 1994-2011 had the character of a natural experiment, allowing checks on what the before/after difference of each area as it was phased in, as against areas where the rollout occurred earlier or later. Moran and Cullinan (2022) looked at data on the ten-year follow-up of 33,722 breast cancer cases. The conclusion was that "while invitation to screening increased detection, it did not significantly decrease the average risk of dying from breast cancer in the population." The authors did however note that "screening may have helped to reduce socioeconomic disparities in late stage breast cancer incidence."

The historical data does, irrespective of questions regarding its current direct relevance, emphasize the importance of tuning breast cancer checks to the risk profile, of finding ways to distinguish progressive from non-progressive cancer, and of avoiding over-treatment. Early diagnosis may allow the use of less invasive forms of treatment. As argued in Esserman and WISDOM Study and Athena Investigators (2017), women want "better, not more screening".

Chapter 5

The uses and limits of observational data

At least in principle, it is relatively straightforward to use regression type methods to make predictions for a set of new data that have been sampled in the same way. What is hard for observational data, harder than is commonly acknowledged, is to give the model coefficients a causal interpretation. For this, it is necessary to have a clear understanding of the processes involved.

- There will be several, perhaps a very large number, of explanatory variables, and an outcome variable.
- The aim is to find a model that will make predictions for new data.
- Note the predictive/descriptive distinction.

- Note the "in sample/out of sample" distinction.
  - But is the "new" a random sample of the old population? (Is the 'target' a random sample of the 'source'?)

5.1 We have a prediction. What are the drivers?

The issues that arise for observational studies do not in general have clear and easy answers. Chapters 20 and 21 of Gelman, Hill, and Vehtari (2020) canvass points that authors of those studies need to address. See also Andrew Gelman's blog.[1] There are no simple answers![2]. All relevant explanatory variables have

[1] https://statmodeling.stat.columbia.edu/2018/11/10/matching-

[2] See also https://mathbabe.org/2011/06/16/the-basics-of-quantitative-modeling/

to be identified, with the manner in which they may be driving predictions then teased out.

Thus, in a comparison between two groups (e..g., in Section 5.4, midwife led versus medical led neonatal care) one variable or factor may be of particular interest, while other variables are used to adjust for differences between the two groups that are at most a secondary focus of interest. Variables that are of secondary interest are commonly referred to as covariates. Regression coefficients can be misleading guides to what is driving predictions if one or more of the relevant covariates is not available or is not properly accounted for. A paradox of the Yule-Simpson type, sometimes referred to as Laird's paradox, has the same potential to mislead.

Little that has been published since Rosenbaum (2002) clarifies greatly the advice that can be given for practical data analysis, beyond what Rosenbaum has to say. Pearl and Mackenzie (2018) ("The Book of Why") offers an interesting assessment. Pearl and his co-author do a good job of highlighting important issues that should be addressed in order to make causality judgments, at the same time overplaying what their methodology can in general achieve. If strictly implemented, the standards are so high that they severely limit what they can in practice achieve. Causality diagrams have a central role. There is a detailed, and insightful, discussion of the history that finally led to the conclusion that smoking causes lung cancer.

5.2 Maternal obesity, and risk of colorectal cancer

Results from a study reported in Murphy et al. (2021) suggest that maternal obesity (30 kg/m2) did increase the risk of colorectal cancer (CRC), by a factor of 2.51.[3] The authors argue that "in utero events are important risk factors for CRC and may contribute to increasing incidence rates in younger adults". They are at the same time careful to acknowledge that, as an observational study, it could not establish cause, and that factors such as diet and microbiome might explain the association. The eating habits of mothers must surely have a large effect on what children eat, both when young and later in life. To what extent might this explain the association.

Obesity is a risk factor for a variety of diseases. Is it obesity that is directly the risk? Or is it dietary and other factors that both increase the risk of obesity

[3] 95% error bounds, as they are termed, run from 1.05 to 6.02, so that the 2.51 risk factor is not very clearly distinguished from 1.

and of associated diseases?

5.3 Cholera deaths in London — 1832 to 1855

6,500 died from cholera in London in 1832. Medical opinion blamed "miasma" or noxious air, associated with the stink from rotting garbage, faeces, and pollution in the Thames. Poor areas had higher rates of cholera, thought to be a result of the more noxious air that arose from crowding and poorer sanitation. Human excreta went into cesspits, with night-soil periodically taken away.[4] In 1842, Edwin Chadwick, in The Sanitary Conditions of the Labouring Population (1842)[5] showed a direct link between poor living conditions, disease and life expectancy.

Under the assumption that miasma from the cesspools and raw sewage pits was the source of infection, the 1848 Nuisances Removal and Diseases Prevention Act[6] was passed that led to the dumping of the raw sewage into the Thames, which was London's main source of drinking water. The 1848-49 epidemic followed shortly after the cesspits were banned. Hassall (1850), in a careful microbiological study, commented:

> … a portion of the inhabitants are made to consume … a portion of their own excrement, and … to pay for the privilege.

By air, or by water — the 1849 epidemic

Farr, who worked as statistician in the UK Registrar General's office, collected data on deaths from cholera in London in the 1849 epidemic. Farr classified districts into three groups thus,
according to the source of the water for most of the householders:

1) Lower Thames, coded as `Battersea`;

2) Sources away from the Thames, coded as `NewRiver`;

3) Further up the Thames than `Batersea`, where the water was less polluted, coded as `Kew`.

[4] See *Cholera epidemics in Victorian London*
https://www.thegazette.co.uk/all-notices/content/100519

[5] https://www.sciencemuseum.org.uk/objects-and-stories/medicine/cholera-victorian-london

[6] Gazette issue 20637

Figure 5.1 summarizes results from a regression analysis that used Farr's data. None of the terms stands out as substantially more important than any other. Higher rates for the poor, where crowded conditions would commonly make it difficult to maintain hygiene, were to be expected.

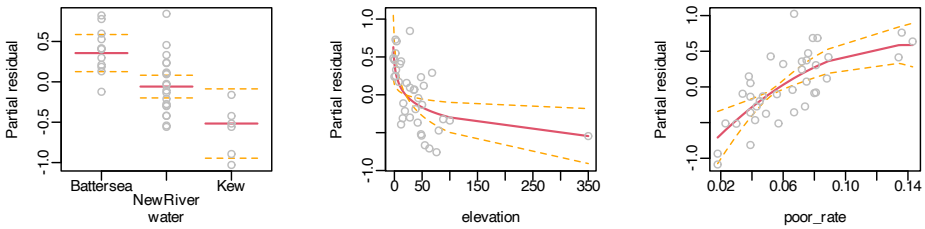

Figure 5.1: Each panel shows, in turn, the estimated contribution of a term in the model relative to the mean contribution from other model terms. Changes in deaths are on a 'log' scale, so that an increase by one unit multiplies the odds of death by close to 2.7, around an overall mean of just over six per 1000.

Snow (1855a) gave examples that he had observed directly, where the likely means of transmission of the infection appeared to be a water source, or poor hygiene. He argued that those living close to the Thames, and especially in the South, were more likely to be getting their water that was contaminated with human excreta. Contaminants had more time to settle in water that was piped up to higher ground.

Farr took Snow's arguments seriously, but in his 1852 report argued that water was primarily important as a vehicle for miasmata. He would later, by the time of an 1866 epidemic when Snow was dead, be one of the waterborne theory's few champions.[7]

A context has to be provided in which to interpret regression results such as those shown in Figure 5.1. Snow's understanding of the contextual information was not, in 1852, sufficiently compelling to persuade other medical specialists. Data from the 1854 epidemic, which allowed a comparison of deaths supplied from a company that continued to get its supply from lower highly polluted Thames water with that from the company that had moved its supply higher up to less polluted water, seems in retrospect to clinch the issue, but did not at the time convince most of the medical profession.[8] The perspective brought by

[7] Eyler (1973)
[8] See Eyler (2004) for further comment.

germ theory had to wait for the work of Pasteur in the late 1850s and Koch in the 1880s.

The 1854 epidemic — a natural experiment

Two water companies — Lambeth, and Southwark and Vauxhall, had been taking water from the same polluted source. An 1852 act required water supply companies to move water intake upriver by 1855. By the time of the 1854 epidemic, Lambeth had moved its intake 22 miles upriver, while the Southwark and Vauxhall intake was unchanged until 1855. Data on the distribution of cholera in the 1854 epidemic then allowed Snow to test the claims made in his 1849 study.

	#Houses	#Deaths	Rate per 10,000
Southwark & Vauxhall	40046	1263	315
Lambeth	26107	98	37
Rest of London	256423	1422	59

> The experiment, too, was on the grandest scale. No fewer than 300,000 people …, from gentlefolks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and, amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity. (Snow 1855b)

Use water from the brewery, and stay healthy!

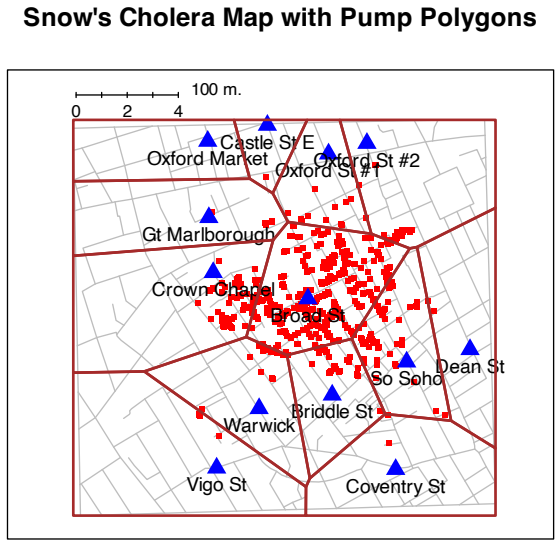

Figure 5.2: Deaths (red dots) and pump locations. Polygons that surround each pump enclose the locations for which that is the nearest pump.

Snow noted that "Within 250 yards of the spot where Cambridge Street joins Broad Street there were upwards of 500 fatal attacks of cholera in 10 days…". By contrast, none of the employees of a local Soho brewery developed cholera. The reason, he judged, was that they drank water from the brewery (which had a different source from the Broad St pump) or just drank beer alone.
Coleman (2019) gives detailed comments on Snow's work.

New Zealand cities had similar issues from the 1840s and 1850s through until the end of the century, arising from failures to install proper drainage systems.[9]

[9]See Christine Dann: 'Sewage, water and waste - Stinking cities', Te Ara - the Encyclopedia of New Zealand, (8 June 2017) https://teara.govt.nz/en/zoomify/24431/dunedin-renamed-stinkapool

5.4 Are there missing explanatory factors?

The (Wernham et al. 2016) study used data from 244,047 singleton term deliveries that occurred between 2008 and 2012 to make the claim that midwife led care, as opposed to medical led care, gave a greater risk of adverse fetal and neonatal outcomes. Notably, the claim was that midwife led care resulted in a lower Apgar score (a measure of infant health immediately after birth) and a greater risk of the imprecisely defined diagnosis of birth asphyxia.

This study was then the basis for exaggerated claims in an article in the October 8-14 2016 issue of the NZ Listener (Chisholm 2016 "Birth Control"). Contrary to what was claimed, the research did not "lob a grenade into the historically war-torn territory of New Zealand's maternity care." Even less did its results warrant the melodramatic claims of "Alarming maternity research" and "Revolution gone wrong" that appeared on the Listener's front cover.

A major issue with the analysis is that it relies on using the NZ Deprivation Index[10] to adjust for socioeconomic differences. This provides a deprivation score for meshblocks, each of around 60–110 people. It estimates the relative socioeconomic deprivation of an area, and does not directly relate to individuals. Deprived areas will often include some individuals with high socioeconomic status. Caesarean section, as a delivery type, may well have been more accessible for those of higher socioeconomic status. For National Women's in Auckland, the elective Caesarean rate at term over 2006-2015 for doctor-led care was 32.8%, as against 7.4% for self employed midwives (Farquhar, McCowan, and Fleming 2016). Effects from fetal alcohol syndrome were not accounted for, nor were direct effects from substance abuse. International data indicates that fetal alcohol syndrome may affect as many as 3% of births.[11]

Studies that are similarly relatively carefully done, but naive in the weight placed on the regression results, are embarrassingly common. There are analysis tools, and associated graphs, that the authors of the study could and should have used to shed light on the likely effectiveness of the adjustments made for differences between the two groups, other than whether the delivery was midwife led or medical led.

[10] https://www.health.govt.nz/publication/nzdep2013-index-deprivation

[11] https://www.health.govt.nz/our-work/diseases-and-conditions/fetal-alcohol-spectrum-disorder

5.5 The uses and traps of rule-based methods

Figure 5.3 shows the distributions of values of six variables that have been selected for use for present illustrative purposes, from an historical dataset (pre 1999, now long past its "use by" date) that has data on 4601 email messages, of which 1813 were identified as spam. In practical use, such datasets have to be continually updated as spammers change their strategies

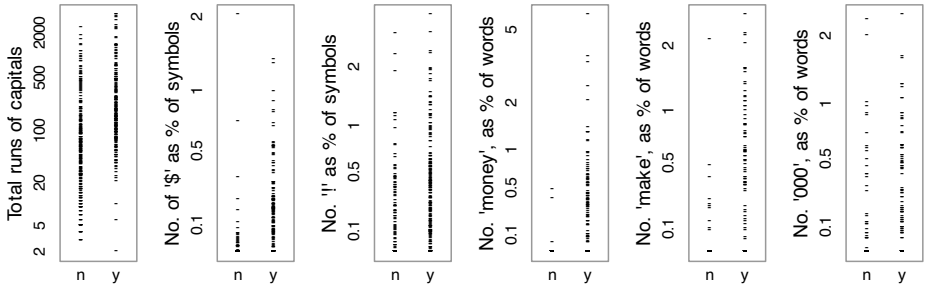

Figure 5.3: Boxplots, showing distribution of variable values in data used to predict email spam

Two types of decision tree approaches will be discussed — the use of individual decision trees, and the random forest approach which generates and uses large numbers of trees in the decision making process.

Figure 5.4 shows a decision tree that has been derived for the spam data.

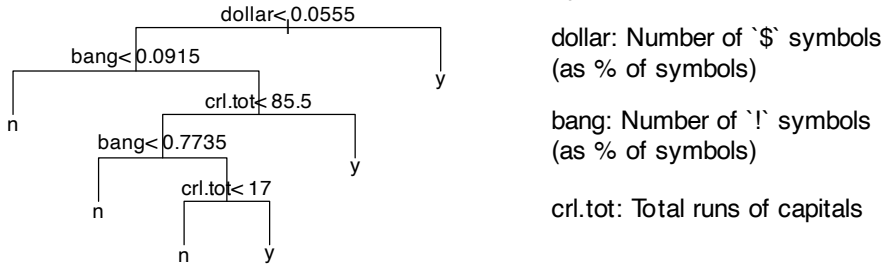

Figure 5.4: Decision tree for spam data. If the condition is satisfied, take the branch to the left. Otherwise, take the branch to the right.

The tree in Figure 5.4 would be too inaccurate for practical use, even suitably updated with new data, but it is easy to follow the decsion tree process.

From trees to forests

"Random forests" improve on decision trees by using samples from the data to create a forest (a "random forest") of trees, then voting between the trees. A downside is that "Random forests" and similar methods operate largely as black boxes. For detection of spam email, this may, as those who deploy the spam detectors have little idea what may be going on in the minds of the spammers, not be too much of an issue. One wants a spam detector that will respond effectively to whatever is thrown at it.

It helps to know the how and why of the algorithms used

Both decison trees and random forests follow an *algorithmic* process. The relatively "black box" nature of the random forest approach places an especially strong burden on the analyst to ensure relevant data have been used, and that the algorithm really is doing its intended task. In her book "Weapons of math destruction", O'Neil (2016) comments:

> … it's not enough to just know how to run a black box algorithm. You actually need to know how and why it works, so that when it does'nt work, you can adjust.

This is too strong. But if one does not know the how and why of how an algorithm works, it is absolutely crucial to be sure that the data used to fit and test the model (the "training" and "test" data) are directly relevant to the task in hand.

Automated systems that can be easily gamed do, however, abound. They are a menace!

Chapter 6

Weighting effects that skew statistics

6.1 Covid-19 deaths — comparing countries

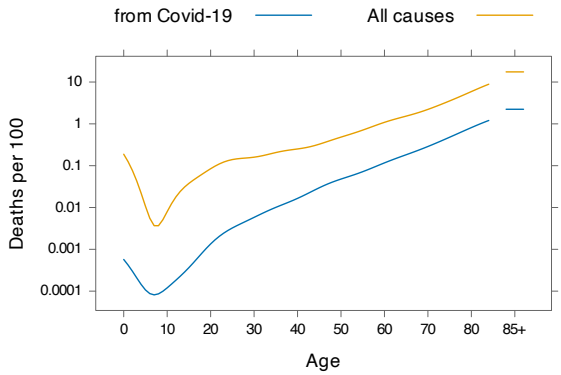

Figure 6.1: US data for proportions who died from Covid-19, and in total, for the 13 months up to 31 January 2021. Vaccination of a substantial part of the population, and the emergence of new variants, will, six months or more later, have substantially changed the pattern of relative risks by age.

Figure 6.1 shows stark differences by age in US Covid-19 death rates. The US death rates per 1000 for under age 65 as opposed to age >= 65, were 0.28, 6.3 US totals for infection rates and for hospital admissions were also impacted

by age structure, but to a lesser extent. More recent Covid-19 variants would present a somewhat different picture.

Countries with a lower proportion of their population aged >=65 would, if the death rates for each of the two groups are similar to US rates, have lower overall death rates. The following compares the overall deaths rates for the US with what might, if these figures carried across, be expected for Kenya and for Italy.

```
                                US Italy Kenya
Percentage 65 or more         16.3  23.3   2.5
Expected deaths per 100,000 126.7 168.5  43.5
Reported deaths per 100,000 126.7 146.0   3.3
```

Between country comparisons are hazardous. The dependence of reported case numbers on testing rates and reporting protocols makes it likely that they will be substantial undercounts, to an extent that varies from country to country.

6.2 University admissions data — Simpson's paradox

Admissions data for University of California Berkeley in 1973 showed a curious anomaly. Overall admission rates strongly favoured males, while in individual departments the rates mostly favoured females. The table shows percent admission rates, with number applying shown in brackets underneath.

	OVERALL	A	B	C	D	E	F
Male	**44.5**	62.1	63	36.9	33.1	27.7	5.9
	(2691)	(825)	(560)	(325)	(417)	(191)	(373)
Female	**30.4**	82.4	68	34.1	34.9	23.9	7
	(1835)	(108)	(25)	(593)	(375)	(393)	(341)

Figure 6.2 provides a graphical summary.

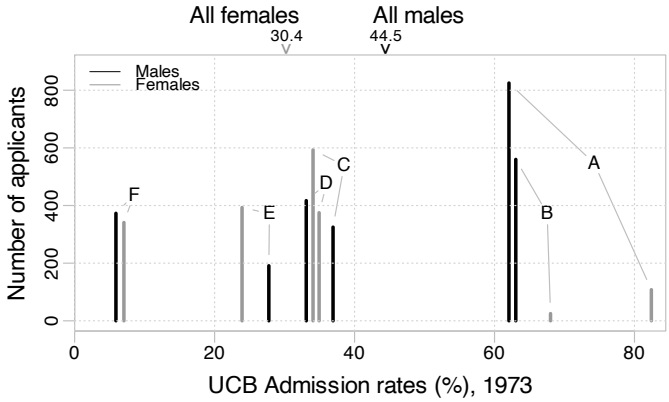

Figure 6.2: UCB admossion data for 1973, for males and females, by department. Department labels range from A to E. Notice that the largest differences in admission rates are for departments A and B, in both cases favouring females.

There are thus three different broad admission patterns. The big differences in admission rates were in departments A (82.4%-62.1%=20.3%) and B (68%-63%=5%), in both cases favouring females. In the other four departments, differences were 3.8% or less, and split very nearly equally between slightly favouring females and slightly favouring males. Also, the relative numbers of males and females applying did not show the same big differences as in departments A and B.

Table 6.1: Comparion of admission rates (percent) and numbers of males and females applying, for departments A, B, and CDEF which combines numbers for departments other than A and B.

	TOTALS	A	B	CDEF
Male	**44.5**	62.1	63	25.5
	(2691)	(825)	(560)	(1306)
Female	**30.4**	82.4	68	26.5
	(1835)	(108)	(25)	(1702)

- The overall male rates are weighted (825+560):(325+(417+191+373 =

1385:1306 or 1.06:1, between an overall AB rate of 62.5%, and the CDEF rate of 25.5%.

- Overall female rates are weighted 133:1702 = 1:15.8 for departments A and B, as against departments CDEF. Overall female rates are, accordingly, strongly weighted towards the 26.5% rate for other departments.

UCB Admissions Data – Another perspective

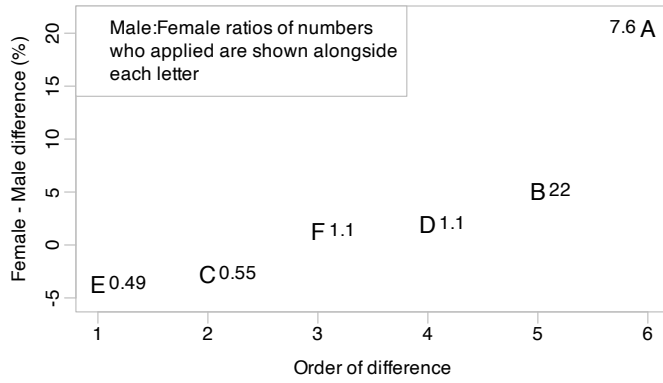

Figure 6.3: UCB admission data for 1973 — another perspective.

In order to understand how the overall imbalance between numbers of males and females arose, it was necessary to break the data down by department. The major driver of the overall imbalance was the low numbers of females, relative to males, applying to department A (and, to a lesser extent, B) where admission rates were highest.

See https://www.youtube.com/watch?v=ZDinnCwP3dg for an animated video that explains the Yule-Simpson paradox.

A note on Lord's paradox

The same sorts of paradoxical effects can be found in regression. The Yule-Simpson paradox may be regarded as a special case of Lord's paradox, described in Lord (1967).[1] Any attempt to attach meaning to regression coefficients can

[1] See also Tu and Gilthorpe (2011), pp.60-71

be highly misleading, unless it is clear that effects of all relevant variables are properly accounted for. It is rarely easy, with observational data, to be sure that this has been done effectively, a point that will be taken up in Chapter 8.

6.3 Comparing unvaccinated with vaccinated

Once high enough vaccination rates have been achieved, a greater number of Covid-19 infections will be found among the vaccinated than among the vaccinated. Suppose that, for a particular age group and for a particular vaccine, the risk rate for severe Covid-19, is 5% for unvaccinated versus 1% for vaccinated, i.e. a relative risk of 5:1. Consider the following scenarios. The final column has the relative numbers, for unvaccinated versus vaccinated with Covid-19:

Unvaccinated	Vaccinated	Relative numbers with severe Covid-19
200	800	10 unvaccinated : 8 among vaccinated
100	900	5 : 9
0	1000	0 : 10

At a certain point, provided the risk ratio is greater than 1.0, the higher number will be from the vaccinated. Of course, the 100% vaccination rate scenario in the final row is unlikely to be realized.

Report found almost 60% of cases were vaccinated

As of August 15 2021, Israeli data showed that 301 of those hospitalized with severe Covid-19 were fully vaccinated, while 214 were not. We need to compare, not the counts, but the proportions of vaccinated and unvaccinated, i.e., look at the percentages that are shown in brackets following the counts.[2] Numbers given are of those aged 12 or more.

	Unvaccinated	Fully vaccinated
Number (aged >=12)	1,302,912	5,634,634
Severe cases	214 (16.4%)	301 (5.3%)

[2]Data are from https://bit.ly/3h1mwXQ, on the website https://www.covid-datascience.com/.

The relative risk is $16.4/5.3 \simeq 3.1$. Because no account has been taken of age differences, this still underestimates the effectiveness of vaccination. A fairer picture is:

	Unvaccinated	Fully vaccinated
Number <50	1,116,834	3,501,118
Severe cases	43 (3.9%)	11 (0.3%)
—-	—-	—-
Number >=50	186,078	2,133,518
Severe cases	171 (91.9%)	290 (13.6%)

The relative risk is $3.9/0.3 \simeq 13$ for those under 50, and $91.9/13.6 \simeq 6.8$ for those over 50, i.e., in both cases much greater than the 3.1 ratio obtained when the split between under 50s and 50 or more is ignored. Even this does not do justice to the data, which needs to be split into smaller age ranges.

6.4 Further illustrative examples

Does Baclofen help in reducing pain?

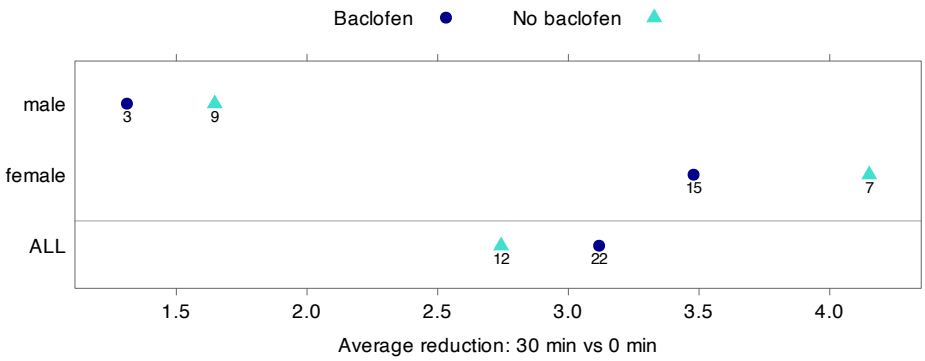

Figure 6.4: Data are pain reduction scores. Subgroup numbers, shown below each point, weight the overall average ("ALL") for baclofen towards the high female average, and for no baclofen slightly towards the low male average.

In work reported in Cohen (1996), researchers were comparing two analgesic treatments, without and with baclofen. When the paper was first submitted for publication, an alert reviewer spotted that some of the treatment groups contained more women than men, and asked whether this might account for the results.

For a fair overall comparison:

- Calculate means for each subgroup separately.
- Overall treatment effect is average of subgroup differences.

The effect of baclofen (reduction in pain score from time 0) is then:

- Females: 3.479 - 4.151 = -0.672 (-ve, therefore an increase)
- Males: 1.311 - 1.647 = -0.336
- Average, male & female = -0.5 × (0.672+0.336) = -0.504

Web page revenue per click

Smith (2014) (p.111) describes an experiment where a US Internet company collected data that compared the effectiveness of two strategies. In the 1-click strategy, an advert appeared on the website's first page. The 2-click strategy required the user to click on a keyword which then led to a page with the advert. The response (total dollar value of purchases) was compared between the two.

The following, with numbers changed to make the comparison relatively simple, shows a scenario that is not unlike that given by Smith:

1-click			2-click		
Revenue	Users	RP1000	Revenue	Users	RP1000
$2500	200	$12.50	$3000	300	$15.00

The 1-click strategy appears to give a better return.

Now, see how the numbers divide up when a split is made between US and international visitors to the site. The numbers (thousands) for the two strategies divide up thus, with the revenue per thousand users given, in each case, in brackets:

	1-click	2-click
US	50	80
	($32)	($30)
Int	150	120
	($6)	($5)

The 1-click strategy clearly gives a better return, both for US and for international users.

The overall figure is dominated by the result for the 150,000 international users (as opposed to 50,000 US) in the 1-click sample. This compares with the much weaker weighting towards international users (120,000 as opposed to 80,000) in the 2-click sample.

In the case that Smith reported, the type of analysis shown was followed up with a randomized experiment (an A/B test), where the probability of assignment to 1-click, as opposed to 2-click, was the same for both classes of user.

6.5 Cricket Bowling Averages

Runs (R), wickets (W) and runs per wicket ($\{RPW\}$)

	1st innings			2nd innings			Overall		
	R	W	*RPW*	R	W	*RPW*	R	W	*RPW*
Bowler A	40	4	*10.0*	240	6	*40.0*	280	10	*28.0*
Bowler B	70	5	*14.0*	50	1	*50.0*	120	6	*20.0*

Fair comparison: Compare runs per wicket ($\{RPW\}$)

	1st innings		2nd innings		Overall	
	RPW	*W*	RPW	*W*	RPW	*W*
Bowler A	10.0	(4)	40.0	(6)	$\frac{10+40}{2} = 25$	(10)
Bowler B	14.0	(5)	50.0	(1)	$\frac{50+14}{2} = 32$	(6)

6.6 Epistatic effects in genetic studies

In population genetics, Simpson's paradox type effects are known as epistasis. Most human societies are genetically heterogeneous. In San Francisco, any gene

that is different between the European and Chinese populations will be found to be associated with the use of chopsticks! If a disease differs in frequency between the European and Chinese populations, then a naive analysis will find an association between that disease and any gene that differs in frequency between the European and Chinese populations.

Such effects are major issues for gene/disease population association studies. It is now common to collect genetic fingerprinting data that should identify major heterogeneity. Providing such differences are accounted for, large effects that show up in large studies are likely to be real. Small effects may well be epistatic.

Chapter 7

Matters of consequence

7.1 The MMR vaccine scandal

The MMR vaccine was developed to for use in preventing measles, mumps, and rubella. Andrew Wakefield was the lead author of a study published in 1998, based on just twelve children, that claimed to find indications of a link between the MMR vaccine and autism. The journalist Brian Deer had a key role in identifying issues with the work, including fraudulent manipulation of the medical evidence.

It emerged that

- Wakefield had multiple undeclared conflicts of interest
- Funding came from a group of lawyers who were interested in possible personal injury lawsuits
- From 9 children said to have regressive autism
  - Only 1 had been diagnosed; 3 had no autism
  - 5 had developmental problems before the vaccine

Wakefield's 1998 claims were widely reported

- Vaccination rates in the UK and Ireland dropped sharply
- The incidence of measles and mumps increased, resulting in deaths and in severe and permanent injuries.

Wakefield was found guilty by the General Medical Council of serious professional misconduct in May 2010 and was struck off the Medical Register.

Following the initial claims in 1998, multiple large epidemiological studies failed to find any link between MMR and autism.

Fact boxes on the Harding site summarize evidence of the effectiveness of the MMR vaccine[1]

7.2 Sally Clark's disturbing cot death story

Sudden Infant Death (SID), also referred to as "cot death", is the name given to the unexplained sudden death of very young children. The story of Sally Clark's unfortunate brush with the law, following the death of a second child, is interesting, disturbing, and educational. Sally's experience highlighted ways in which the UK legal system needed to take on board issues that affect the use of medical evidence — issues of a type that can be important in medical research.

Paediatrician Sir Roy Meadow had argued in his 1997 book **ABC of Child Abuse** that "unless proved otherwise, one cot death is tragic, two is suspicious and three is murder." Was this an example of "Too hard! Try something easier, and wrong!" Or was it the triumph of assumed "knowledge" over hard evidence?

Meadow's role as an expert witness for the prosecution in several trials played a crucial part in wrongful convictions for murder. The case that attracted greatest attention is that of Sally Clarke, both of whose children were "cot death" victims. Following the death of the second child, Meadow gave evidence at her 1999 trial, and appeal in 2000. Sally Clark was finally acquitted in 2003.

- Meadow gave 1 in 8,500 as cot death rate in affluent non-smoking families
  - Squared 8,500 to get odds of 73,000,000:1 against for two deaths.
  - Meadow assumed, wrongly, that the probability of a death from natural causes was the same in all families.
  - A first "cot death" is, in some families at least, evidence of a greater proneness to death from natural causes.
- Royal Statistical Society press release: "Figure has no statistical basis"
  - 2000 appeal judges: The figure was a "sideshow" that would not have influenced the jury's decision.
  - The appeal judges' statement was described by a leading QC, not involved in the case as: "a breathtakingly intellectually dishonest judgment."
- 2003: It emerged that 2nd death was from bacterial infection

[1] https://www.hardingcenter.de/de/search/node?keys=mmr

  - In a second appeal, Sally Clark was freed.
- 2005: Meadow was struck from medical register
  - 2006: reinstated by appeal court — misconduct fell short of "serious"!

Meadow was in effect assuming, without evidence, the absence of family-specific genetic or social factors that make cot deaths more likely in some families than in others. Why was Meadow's assumption of independence not challenged in the 1999 trial? Issues of whether or not different pieces of evidence are independent are surely crucial to assessing the total weight of evidence. Meadow had only enough understanding of probability to be dangerous.

Sally Clark was finally acquitted on her second appeal in 2003, with her sense of well-being damaged beyond repair. The forensic evidence had been weak. The web page https://plus.maths.org/content/os/issue21/features/clark/index has a helpful summary of the statistical issues. It quotes a study that suggests that the probability of a second cot death in the same family is somewhere between 1 in 60 and 1 in 130. Even after this adjustment, the probability of death from natural causes of two children in the same family is low. But so also is the probability that an apparently caring mother from an affluent middle class family, with no history of abuse, will murder two of her own children. Those are the two probabilities that must be compared. Anyone who plans to work as a criminal lawyer ought to understand this crucial point. In a large population, there will from time to time be two deaths from natural causes in the one family.

A small number of appeals were subsequently launched against other convictions where evidence of the same type had been presented, most of them successful. A further consequence was that the law was changed such that no person could be convicted on the basis of expert testimony alone.

The Watkins (2000) article "Conviction by mathematical error?: Doctors and lawyers should get probability theory right" provides a trenchant critique of the perils of allowing non-statisticians to present unsound statistical arguments, with no effective challenge.

> Guidelines for using probability theory in criminal cases are urgently needed. The basic principles are not difficult to understand, and judges could be trained to recognise and rule out the kind of misunderstanding that arose in this case.

Watkins, who was at the time Director of Public Health for Stockport in the UK, argued that the calculation of the relevant probability should have had no regard to the probability of an initial cot death. It surely has some relevance.

The basic principles are perhaps more difficult than Watkins was willing to allow.

7.3 The Reinhart and Rogoff saga

Figure 7.1 plots data, for 1946 – 2009 from 20 "advanced" countries, that underpinned the 2010 paper "*Growth in Time of Debt*" by the two Harvard economic historians Reinhart and Rogoff

$$RR$$

.

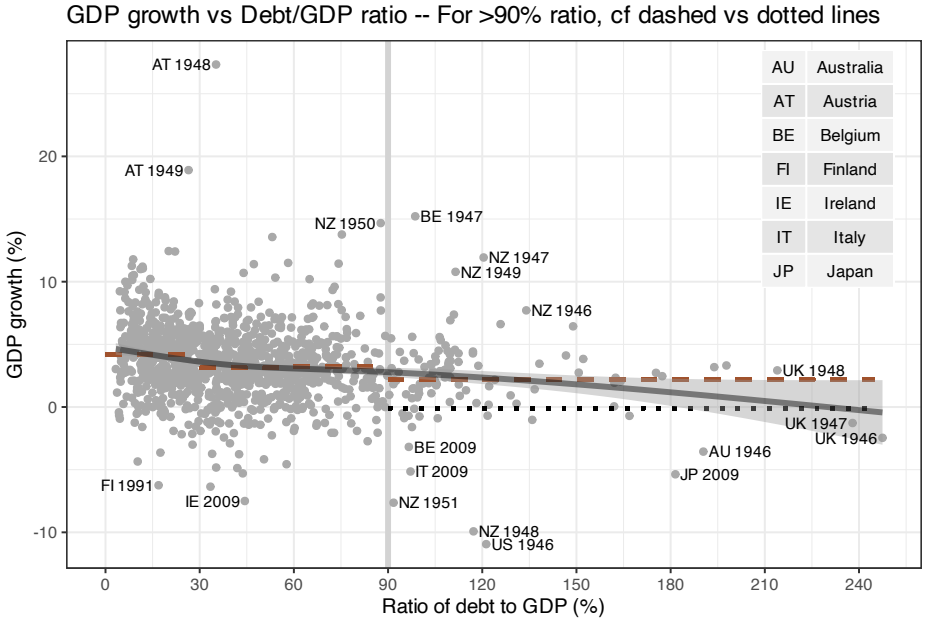

Figure 7.1: Dashed horizintal lines show means by Debt/GDP category, for 20 'advanced' countries, for the years 1946 — 2009. (Data are missing for some countries in some years.) RR's mean (dots) for > 90% Debt/GDP was from 10 only of the 20. The smooth gray curve treats points as independent.

The paper (Reinhart and Rogoff 2010) has been widely quoted in support of economic austerity programs internationally. There was a huge stir, in the media

and on the blogosphere, when graduate student Herndon found and published details of coding and other errors in the results that RR had presented.[2]

As well as coding errors, Hendon identified selective exclusion of available data, and unconventional weighting of summary statistics. There was no consideration that the relationship studied has varied substantially by country and over time. Half of the 20 countries had missing data for 5 or more of the years, with the largest number missing in the years 1946 to 1949.

In response to Herndon, Ash, and Pollin (2014) and other critics, Reinhart and Rogoff accepted that coding errors had led to the omission of several countries, but pushed back against other criticisms. Their revised analysis addresses only the most egregious errors in their work. Among other issues, their insistence on treating each data point for each country as an independent piece of evidence makes no sense. The smooth curve fitted in Figure 7.1 may be regarded as an average over the ten countries, but as Figure 7.2 shows, with huge country to country variation.

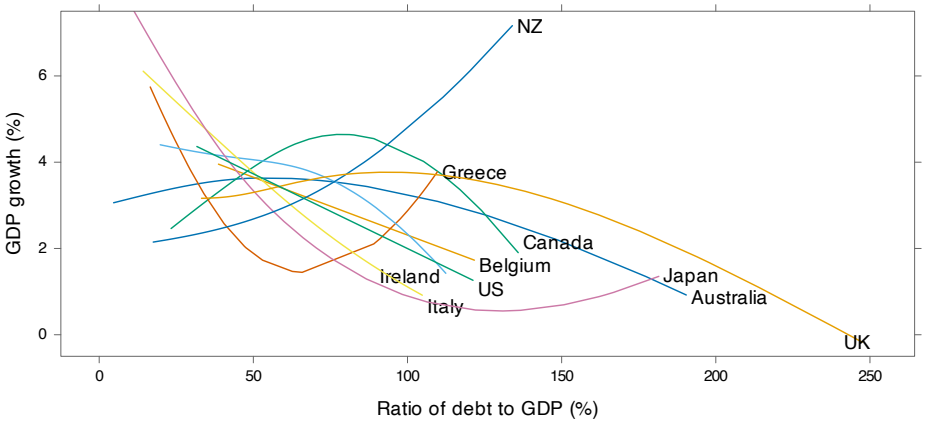

Figure 7.2: Smooths have been fitted for each of the 10 countries for which debt to GDP ratios were in some years greater than 90%. There is no consistent pattern, as there should be if RR's claim is to hold up.

[2] http://www.peri.umass.edu/fileadmin/pdf/working_papers/working_papers_301-350/WP322.pdf

Is there a pattern across countries?

There were just 10 countries where debt to GDP ratios were greater than 90% for one or more years. For 9 of those countries, the average of their GDP growth in the years at issue was on average positive relative to the previous year. The average percentage growth over all 10 countries, weighted according to number of years, was 2.168.

There are further serious issues of interpretation

- Does GDP drive debt/GDP ratio, or is it the other way round?
  - Or does a third factor drive both?
- Is the effect immediate, or on future economic performance

Smith (p.64) refers to work indicating that economic performance is more closely correlated with economic growth in the past than with future growth.

Parting comments

Herndon, Ash, and Pollin (2014) comment

"... RR's findings have served as an intellectual bulwark in support of austerity politics. The fact that RR's findings are wrong should therefore lead us to reassess the austerity agenda itself in both Europe and the United States."

The saga emphasizes the importance of working with reproducible code, rather than with spreadsheet calculations. The errors in RR's calculations were from one perspective fortunate. Once highlighted, the errors drew critical attention to the paper, and to the serious flaws in the analysis.

7.4 What do malaria drugs do to Covid-19 patients?[3]

Thirteen days after it was published on May 20 2020, three of the four authors withdrew a paper that claimed to find that malaria drugs, when used experimentally with patients with Covid-19, led to around 30% excess deaths. Irrespective of the problems with the data that will be noted shortly, serious flaws in the analysis ought to have attracted the attention of referees. There was inadequate adjustment for known and measured confounders (disease severity, temporal effects, site effects, dose used).

[3] Lancet, May 2020, https://bit.ly/3xqncMt

The study claimed to be based on data from 96,032 hospitalized COVID-19 patients from six continents, of which 66% were from North America. Very soon after it appeared, the article attracted critical attention, with a number of critics joining together to submit the Watson et al. (2020) letter to Lancet.

The sources from which the data had been obtained could not be verified, data that claimed to be from just five Australian sources had more cases than the total of Australian government figures, and similarly for Australian deaths, there were implausibly small reported variances in baseline variables, mean daily doses of hydroxychloroquine that were 100 mg higher than US FDA recommendations.

Randomized trials designed to test the effectiveness of the drugs, and that were in progress at the time when the paper appeared, were temporarily halted. The eventual conclusion was that the drugs did not improve medical outcomes. There was some evidence that hydroxychloroquine could have adverse effects.

With current web-based technology, randomized controlled trials can be planned and carried out and yield definitive answers, in much the same time as it would take to collect and analyze the data that are required for an observational study whose conclusions can be, at best, suggestive. Data confidentiality issues are easier to handle in the context of an RCT.

7.5 A simplistic use of publicly available data

A June 2021 paper in *Vaccines*, titled "The Safety of COVID-19 Vaccinations -— We Should Rethink the Policy."[4] massively over-stated vaccine risk. The paper claimed that "For three deaths prevented by vaccination we have to accept two inflicted by vaccination."

- Deaths were from any cause, post-vaccination, reported both by professionals and the public
  (Such data has to be used with a 'baseline' comparison)
- Vaccine benefits extend far beyond 6 weeks of Israeli study
- The paper focused on immediate risk to individual, not community.
- The way that any risk from the vaccine balances out against risk of death from Covid-19 will vary depending on the the age structure of the population, on proportion immunized within each age group, and on the age and health of the individual.

[4] *Vaccines*, June 2021, bit.ly/3dTg1oh

Deaths that could be verified as from the Pfizer vaccine have been, with the possible exception of frail elderly people, extremely rare. See US CDC report on risk (for Pfizer, anaphylaxis)[5], and Helen Petousis-Harris' commentary on risks.[6]

[5] www.cdc.gov/coronavirus/2019-ncov/vaccines/safety/adverse-events.html

[6] sciblogs.co.nz/diplomaticimmunity/2021/04/15/covid-vaccines-and-blood-clots-what-is-this-about/

Chapter 8

Regression and Correlation

When two variables show a relationship, they are said to be correlated. Regression and correlation offer alternative, and complementary, perspectives on the relationship.[1] Nonsense correlations that arise where the third variable is time provide simple examples.[2]

8.1 Correlation is not causation

Variable A may cause variable B. Or variable B may cause variable A. Or both A and B may be caused (or driven) by a third variable C.

The following are examples where causation likely goes in the other direction, or where a third variable is likely to be involved. Such examples help highlight how correlation can and cannot reasonably be interpreted.

1. Children of parents who try to control eating are more likely to be overweight.
2. Ice cream consumption & polio were closely correlated in the 1950s.
    - Summer was when the virus thrived.

Cases where the third variable is time, as in Figure 8.1, are a fruitful source

[1] The discussion will avoid the technicalities of alternative ways to measure correlation. The particular correlation measure that is used in the examples that follow is the product-moment or Pearson correlation, which relates directly to linear (straight line) regression.

[2] Yule-Simpson type effects, discussed in Section 6, are important in a regression context also.

of examples of spurious correlations.[3] Is there a third factor, or is this just a chance relationship?

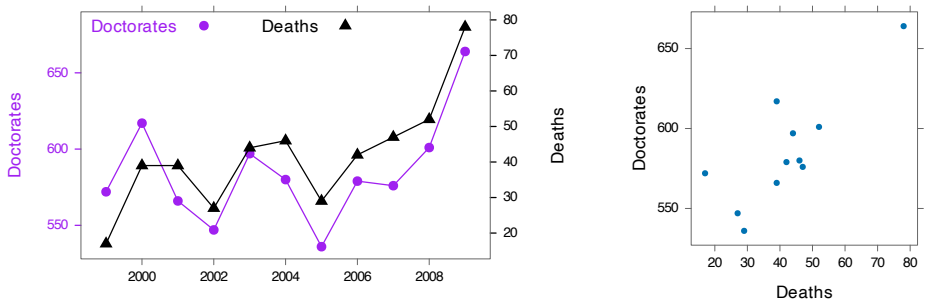

Figure 8.1: Sociology PhDs awarded (from US National Science Foundation data) vs Deaths from Anticoagulants. Notice that 'Doctorates' and 'Deaths' show a very similar pattern of change from year to year.

Another example is that, over 1997 to 2009, US Sociology PhDs awarded correlated strongly with worldwide non-commercial space launches. These correlations would seem to be the result of chance. Look at enough variable pairs, and such correlatiosn will sometimes appear.

8.2 Regression to the mean

Tall fathers are likely to have tall sons, but shorter than themselves. Tall sons are likely to have tall fathers, but shorter than themselves. The data shown in Figure 8.2 are from Pearson and Lee (1903). The correlation between son's height and father's height is 0.5.[4]

Notice that the points that are plotted show a a symmetrical elliptical shaped scatter about the mean (shown with a large solid dot in Panel A). This type of scatter is, strictly, required for uses of the correlation that will now be discussed.

[3] Figure 8.1 is one of many such examples that are available from http://www.tylervigen.com/spurious-correlations.

[4] Kahneman argues, perhaps too simplistically that as height is mainly due to genetic factors, and fathers share half of their genes with their sons, this is to be expected.

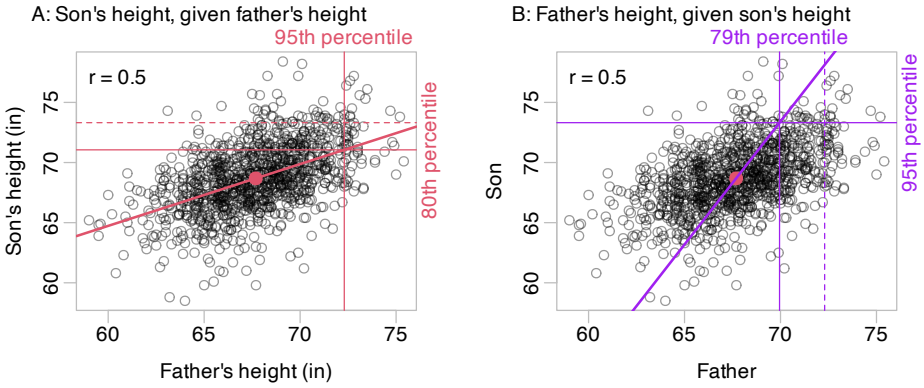

Figure 8.2: Tall fathers are likely to have tall sons, but shorter than themselves. Tall sons are likely to have tall fathers, but shorter than themselves.

Consider a father whose height is 72 inches, which is the 95th percentile of heights for fathers. What is a best guess for the height of a son? One can read the predicted value off from the graph (the solid horizontal line in Panel A). Or use the regression equation. Or, reason thus:

- If the correlation between father's height and son's height were 0, the best guess would be the mean for son's, i.e., 68.7 inches
- If the correlation were 1, the son's height would be the 95th percentile of heights for sons, i.e., 73.3 inches.
- But, as the correlation is 0.5, the expected height is 68.7 + 0.5×(73.3-68.7) = 71.0 inches, i.e., start at the mean and move 0.5 of the distance up to the 95th percentile.

Now consider a son whose height is 73.3 inches (the 95th percentile for sons). The argument now goes:

- The best estimate of the father's height is 67.7 + 0.5×(73.3-67.7)=2.8 inches, i.e., start at the mean for fathers and move 0.5 of the distance up to the 95th percentile.

Galton's 1886 data, which predates Pearson's data, shows a 0.46 correlation between child height and the average of the parent height.

8.3 NBA player points — correlations decline over time

In Figure 8.3, Panel A shows total points for 2016-2017 versus 1 year earlier, for players who competed in both seasons. The correlation is 0.83. Panel B is for 2016-2017 versus 5 years earlier, with the correlation now reducing to 0.41.

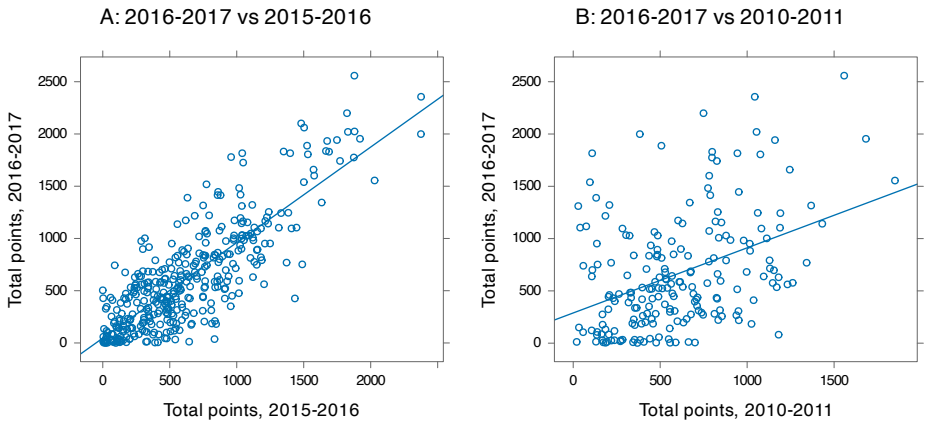

Figure 8.3: As time progresses, correlation decreases, and regression to the mean increases. For Panel A, the correlation is 0.83, while in Panel B it is 0.41.

The scatter of points increases as values increase, on both axes. Calculations of the type given in the previous section, based on the usual correlation measure, while giving more approximate results, are adequate for present purposes.

8.4 Secrist's "The Triumph of Mediocrity in Business"

Horace Secrist's 1933 book The Triumph of Mediocrity in Business" was based on annual data for 1920 to 1930. Secrist took 73 different industries, in each case examined the ratios

Profits:sales; Profits:assets; Expenses:sales; Expenses:assets

For each industry in 1920: he then split firms into 4 quartiles: top 25%, 2nd highest 25%, 2nd lowest 25%, lowest 25%.

- Took average for each statistic, for each quartile, for each year.

- Surprise, surprise, the best went, on average, down …

  Complete freedom to enter trade and the continuance of competition mean the perpetuation of mediocrity. … neither superiority or inferiority will tend to persist. Rather, mediocrity tends to become the rule. (Secrist 1933)

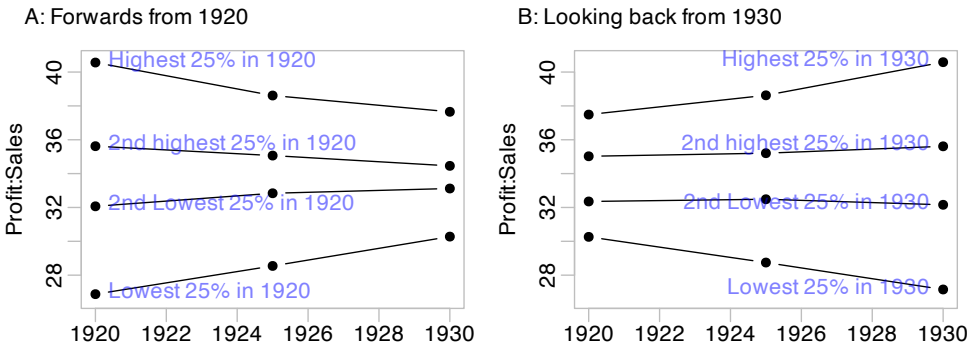

Figure 8.4: Secrist's data showed a correlation of 0.5 between time intervals five years apart. Panel A uses shows means of simulations, starting with the four performance quartiles in 1920 and looking ahead. Panel B starts with the equivalent quartiles in 1930, and looks back.

Secrist was seeing regression to the mean. Figure 8.4 makes the point that if one takes the four quartiles in 1930 and looks back to 1920, in each case there is a regression back to the mean. Given a correlation of 0.7 between time intervals five years apart, The absolute difference from the mean moves from 8 to 0.7× 8 (= 5.6) to 0.7× 5.6 (= 3.92), whether one moves by two successive five year intervals forward in time, or back in time.

"Do old fallacies ever die?"

Smith (2014) gives references to work by prominent economists in the past half-century that had quoted Secrist approvingly or repeated his error.

- 1980s investment textbook: "Ultimately, economic forces will force the convergence of the profitability and growth rates of different firms." This was backed up with a 1980/1966 Secrist type comparison.

- 2000: (Journal article) "… profitability is mean-reverting within as well as across industries. Other firms eventually mimic products and technologies that produce above normal profitability …"
- Wainer (2000) cites other examples.

Decathlon performances in 2006

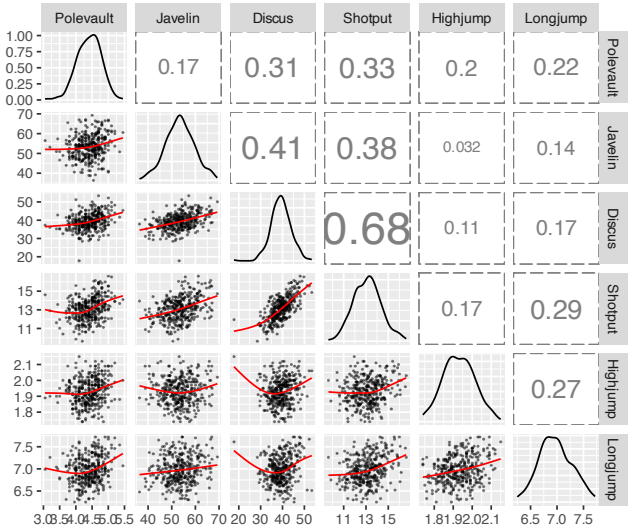

Figure 8.5: Between event correlations for top performances in six of the ten decathlon events in 2006. Points that are plotted, and correlations, are for times or distances achieved.

Figure 8.5 shows between event correlations for top performances (6800 points and over) for six events in the 2006 decathlon.[5] Note the 0.032 correlation of javelin with high jump. Performance in one of these two sports was not a useful indicator of what to expect in the other.

The correlation between shot put distances and long jump distances is shown as a more informative 0.29. If we find that an athlete has achieved a distance of

[5]The dataset `Decathlon` in the R package `GDAdata` has data for the 21-year period after new rules were introduced in 1985. See also the Estonian website http://www.decathlon2000.com/

14.9 meters in the shot put, which is at the 93^{th} percentile across athletes as a whole (7% did better), a best estimate for the long jump can be obtained thus:

- A 14.9 meter put is at the 93^{th} centile (7% will do better)
- The long jump mean is 6.97, with the 93% mark = 7.47
  - The difference from the mean is 7.47-6.97 = 0.5
- The estimate for the long jump is then 6.97+0.29 × 0.5 = 7.12

8.5 Moderating predictive assessments

Moderating sales estimates

You are the sales forecaster for a department store chain. All stores are similar in size and merchandise offered, but random factors affect sales in any year. Overall sales are expected to increase by 10% from 2020 to 2021. Sales in 2020, with the expected total and mean for 2021 are, in millions of dollars:

Store	2020	2021
1	10	—
2	23	—
3	18	—
4	29	—
TOTAL	80 t each of the	

stores in 2021. The mean sales amount in 2021 is predicted to be 22,000,000 dollars? With a correlation of 0.4, the predicted sales for the individual stores are obtained thus:

Store	2020	Subtract 20	Xply by 0.4, add to 22	Predicted sales
1	10	-10	22-4	18
2	23	+3	22+1.2	23.2
3	18	-2	22-0.8	21.2
4	29	+9	29+3.6	32.6
MEAN	20	0	22	22

In the real world of 2020 and 2021, the Covid-19 pandemic makes all such predictions hazardous!

Choosing from job applicants

Correlation between presentation & performance is likely to be lower for the less well-known. In both cases performance is likely, relative to presentation, to move in closer to the mean. For less well-known candidates, the shift towards the mean is likely to be greater.

Kahneman's comments on regression to the mean

> "Extreme predictions and a willingness to predict rare events from weak evidence are both manifestations of System 1. …"
>
> "Regression to the mean is also a problem for System 2. The very idea … is alien and difficult to communicate and comprehend. This is a case where System 2 requires special training."
>
> "We intuitively want to match predictions to the evidence."
>
> "We will not learn to understand regression from experience."

Regression to the mean in verse

https://www.youtube.com/watch?v=sxMllckUWaw

8.6 Time per unit distance for hillraces

Regression coefficients may differ greatly depending on what adjustments are made for other variables. This is important for attaching meaning to a coefficient. For the hillrace data that will now be considered, it is relatively easy to tease out the role of the explanatory variables that have been included in one or alternative versions of the regression equation. Especially where there are three or more explanatory variables, with the manner in which they should enter into the regression equation is unclear, the effect of an individual variable that is of interest can be difficult or impossible to tease out.

The hillrace dataset has record times for 23 Northern Ireland Mountain Running Association hillraces, as given in the 2007 calendar. In the models fitted and graphs shown that follow, the distance measure is `Dist` (distance converted to kilometers), the climb measure is `Climb` (vertical distance between lowest and highest point, in meters), and the time measure is `Time` (in minutes).

How does time per unit distance (`timePerKm`) vary with distance. We will fit two equations, both with $y =$ `timePerKm`.

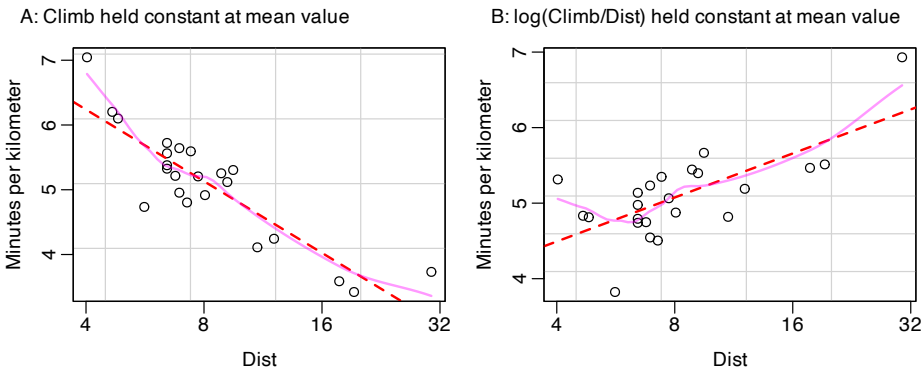

Figure 8.6: Variation in time per unit distance with distance. Panel A shows the pattern of change when 'log(Climb)' is held constant at its mean value, while Panel B shows the pattern of change when 'log(Climb/Dist)' is held constant at its mean value.

Figure 8.6A shows the dependence of `timePerKm` depends on `log(Dist)`, when `log(Climb)` is held at its mean value. Use of `log(Dist)` rather than `Dist` means that distance on the x-axis from 2 to 4 (km) is the same as from 4 to 8, or from 8 to 16, or from 16 to 32, i.e., equal distances correspond to equal multiplicative changes. The equation that is plotted is

$$\texttt{timePerKm} = 8.5 \text{ -}1.6 \times \texttt{log(Dist)}$$

Figure 8.6B shows the dependence of `timePerKm` depends on `log(Dist)`, when `log(Climb/Dist)` is held at its mean value. The equation that is plotted is

$$\texttt{timePerKm} = 3.33 + 0.84 \times \texttt{log(Dist)}$$

In Panel A, time per kilometer decreases quite sharply as distance increases. This happens because the ratio of `Climb` to `Dist` decreases if `Climb` is held constant while `Dist` increases, i.e., longer distance races involve gentler ascents and descents.

Panel B shows what happens when `Climb/Dist` is held constant, i.e., we are comparing races with the same ratio of `Climb` to `Dist`. As expected, time per

kilometer does then decrease as distance increases.

8.7 Model that do not correctly fit the data readily mislead

Are herricanes more dangerous than himmicanes?

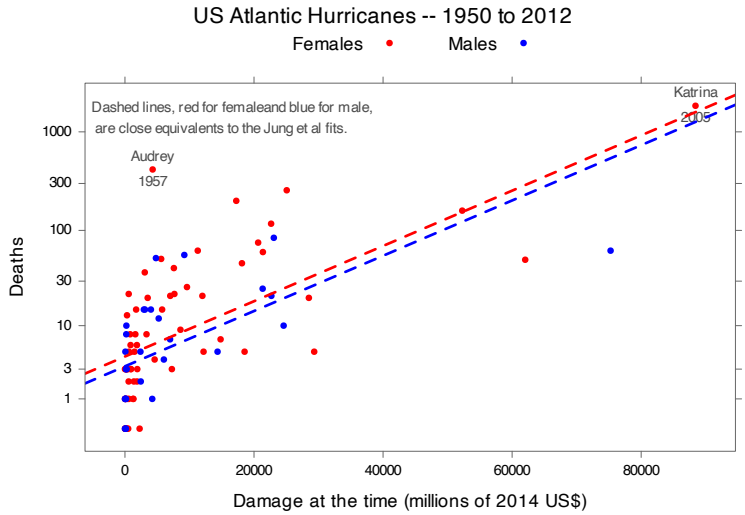

Figure 8.7: Deaths versus damage estimate in US dollars. The red (for female) and blue (for male) dashed lines are close equivalents of Jung at al's fit to the data. The y-axis uses a scale of equal relative numbers of deaths, while the x-axis uses a scale of equal dollar damage costs.

The United States National Hurricane Center began formally naming hurricanes in 1950, a task now under control of the World Meteorological Organization. Female names were used for Atlantic hurricanes from 1953 to 1978, with a mix of male and female names used from 1979 onwards.

In a paper titled "Female hurricanes are deadlier than male hurricanes", Jung et al. (2014) used data for 94 Atlantic hurricanes that made landfall in the United States during 1950-2012 to argue that death rates from those with female names were overall higher than for those with male names. The suggestion

was that where names were female, authorities took the risk less seriously. The paper attracted wide interest on the blogosphere, with female hurricanes jokingly called herricanes and males called himmicanes.

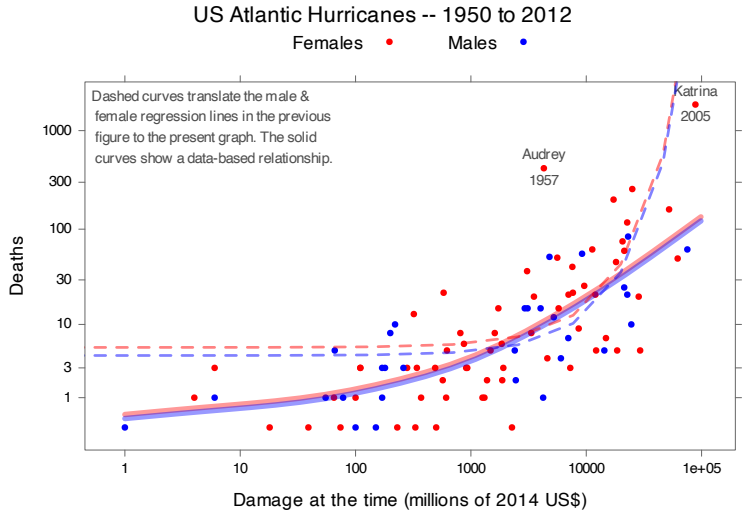

Figure 8.8: Deaths versus damage estimate in US dollars, with logarithmic scales on both axes. Separate fitted lines for male and female hurricanes cannot be distinguished. Jung et al used a logarithmic scale on the vertical axis only, which on this graph leads to the dashed curves.

The separate dotted lines in Figure 8.7, red for female and blue for male, are a close equivalent to the authors' fit to the data. Notice the use of a relative (numbers of deaths) scale on the y-axis, and a dollar scale on the x-axis. An unfortunate consequence of the use of a linear dollar scale on the x-axis is that the slopes of the lines are strongly influenced by the final four points at the upper end of the scale. Why did the authors not use, at least as a starting point, the same relative scale on both axes, as in Figure 8.8?

As well as using a relative scale on the x-axis, Figure 8.8 uses a methodology that allows the data to determine the form of the response. Deaths do on average increase more at a higher rate than the damage measure, but not at the rate suggested by the dashed curves. There is now no evident difference between the two curves.

Jung et al omitted `Audrey` (in 1957) and `Katrina` (in 2005), as outliers. These are included in Figures 8.7 and 8.8, with the curves fitted using a "robust" fitting method that is relatively insensitive to outliers. Other differences between the Jung et al analysis, and the analyses reflected in Figures 8.7 and 8.8 are documented in Note 1 , on p. 87

Historical speed of light estimates — is there a pattern?

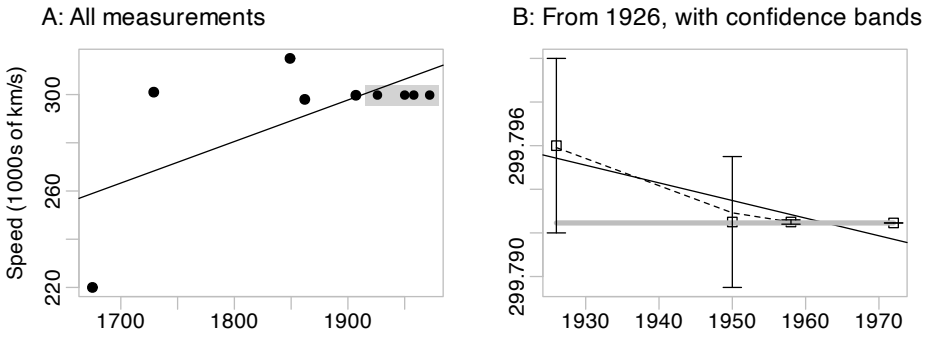

Figure 8.9: Successive speed of light estimates. Panel B limits attention to measurements made in 1926 and later. The line was fitted with no adjustment for the very different error estimates. The dashed curve, which incorporates such adjustments, is statistically indistinguishable from the thick gray horizontal line.

Creationist Barry Setterfield has argued that a reduction over time in the speed of light has led the passage of time to slow down, relative to the remote past, so that the universe is thousands rather than billions of years old. His arguments rely on making various adjustments to figures obtained historically, selecting what he regarded as the most reliable data, and then fitting a curve.

Setterfield tells a story that, while a little different from that of the line in Panel A of Figure 8.9, makes equally little sense. The right panel is limited to the points from 1926 and on, marked off with the gray background on the left panel.[6]

For the measurements from 1862 onward, estimates of accuracy are available. Until 1950, each new estimate lay outside the bounds for the previous estimate,

[6]Data are from https://en.wikipedia.org/wiki/Speed_of_light

indicating that these were underestimates. Even if one were to accept Setterfield's manipulation of the data, it makes no sense at all to fit either lines such as are shown, or curves, to data values which have such very different accuracies.

Even if one were to accept Setterfield's manipulation of the data, it makes no sense at all to fit either lines such as are shown, or curves, to data values which have such very different accuracies as those shown in the graphs. For the measurements from 1862 onwards, estimates of accuracy are available. Until 1950, each new estimate lay outside the bounds for the previous estimate, indicating that these were underestimates.

8.7.1 Global mean temperature trends

Figure 8.10 plots global air and sea surface temperature anomaly data against year. Anomalies, in hundredths of a degree centigrade, are differences from the 1951-1980 global average. The grey curve plots the average anomaly up to that point in time.

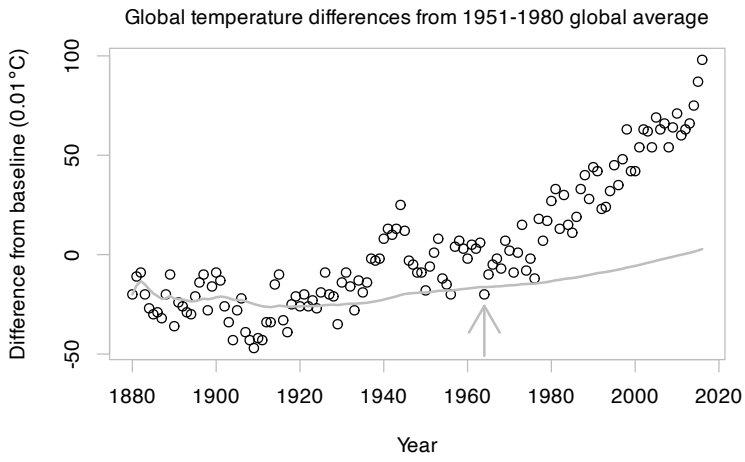

Figure 8.10: Anomalies (differences) in hundredths of a degree centigrade from global average temperatures over 1951-1980, plotted against year. The gray curve shows, for each year, the average anomaly up to that point in time. The last year in which this lay below the gray line was 1962.

Observe that 1964 was the last year in which the global temperature fell below

the average to that time. For the 52 subsequent years (from 1965 to 2016 inclusive), the global average was above the average up to that date. Under the (false) assumption that global temperature is varying randomly (and therefore independently) about a common mean, the probability of this happening is $2^{-40} = 9.1 \times 10^{-13}$. A variation of this argument came from a speaker on the Australian ABC Science Show on April 3 2011.

Under any model that accounts for what are now fairly well understood patterns of correlation over time, the probability, while very small, is not that small! Arguments that overstate the case for what is now a well-established pattern of change are unhelpful

It is likewise nonsensical to fit a line to the cherry-picked years 1998-2008, where the trend was relatively flat. Year to year temperatures are correlated.

Chapter 9

Critiquing scientific claims

To be credible, scientific claims must be able to survive informed criticism. The nature of the critique that is needed will vary, depending on what is in question. It will vary, at a broader level, depending on whether what is in question is

- Measurements
  - e.g., as in astronomy, distances to other planets, stars, and galaxies
- or the results of an experiment
  - e.g., showing that plants grow better when compost is added to a nutrient deficient soil
- or a theory, such as Newton's law of gravity or laws of motion.
  - A theory is a model that is designed to describe natural phenomena. A theory in conceptual, relying on entities such as electrons and waves that lie outside of normal human experience, where a law is a mathematical description of observable phenomena.
- Large parts of science, and scientific applications, rely on models that in turn build on scientific theories/laws.
  - Epidemiological models have been crucial to informing New Zealand's response to the Covid-19 pandemic.
- Climate models are an example of very complex models. These rely on
  - The basic laws of physics, fluid motion, and chemistry
  - Knowledge of the way that atmosphere, oceans, land surface, ice, and solar radiation interact to change climate.
  - Computer simulation to build in the effects of areas of uncertainty.

Measurements and experimental results must be replicable — i.e., another ex-

perimenter must be able to repeat the experiment and obtain the same results. Theories must be able to make successful predictions.

In areas where results depend on the sharing of data and skills between different scientists and groups of scientists, the critique that authors provide to the work of their fellow authors will commonly ensure that what is submitted for peer review is soundly based.

For experimental studies that are designed to stand on their own, the past decade has seen the emergence of concerning evidence that a large amount of published work is, when put to the test, not replicable. The steps needed to implement change are well understood. The slow pace of reform is disappointing.

More generally, uncritical and faulty statistical analyses, such as have been documented earlier in this book, are a cause for concern. Too often, model fitting becomes a ritual, without the use of standard types of diagnostic checks that would have demonstrated that the model was faulty.
Stark and Saltelli (2018) comment:

> The mechanical, ritualistic application of statistics is contributing to a crisis in science. Education, software and peer review have encouraged poor practice–and it is time for statisticians to fight back. . . . The problem is one of cargo-cult statistics - the ritualistic miming of statistics rather than conscientious practice. This has become the norm in many disciplines, reinforced and abetted by statistical education, statistical software, and editorial policies.

It is not just statisticians who should be fighting back, but all scientists who care about the public credibility of scientific processes.

9.1 What results can be trusted?

Scientific processes work best when claims made by one scientist or group of scientists attract widespread interest and critique from a wider group of scientists who understand the work well enough to provide informed and incisive criticism. This can be an effective way to identify claims that have no sound basis. Examples are the May 2020 Lancet and New England Journal of Medicine studies, claiming to be based on observational data, arguing that use of the drug hydroxychloroquine as a treatment for Covid-19 was increasing patient deaths. Issues with these papers were quickly identified because they made claims that bore on an issue of major concern, and attracted attention from readers who

carefully scrutinized their detailed statements. They were quickly retracted. How much that has no sound basis does that attract such attention, and is never challenged?

Heavy reliance on the sharing of data and skills, and full use of the benefits that modern technology has to offer, have been vital to progress in such areas as earthquake science, the study of viruses and vaccines, modelling of epidemics, and climate science. This sharing of data and skills, and use of modern technology, also helps in the critique of what has been published earlier. Areas where there has not been the same impetus for change are much more susceptible to the damage that arises from systems for funding and publishing science that encourage the formal publication of what would better be treated as preliminary results — a first stab at an answer. Publication of experimental results should not be a once-for-all event, but a staged process that moves from "this looks promising" to "has been independently replicated", and to post-publication critique.

Publication does not of itself validate scientific claims, Rather, as stated in Popper (1963)

> Observations or experiments can be accepted as supporting a theory (or a hypothesis, or a scientific assertion) only if these observations or experiments are severe tests of the theory – or in other words, only if they result from serious attempts to refute the theory.

The demand that experimental results are shown to be replicable is central to effective scientific processes. As Fisher (1937) wrote

> No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon . . . In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure.

Sources of failure

Fraud, though uncommon, happens more often than one might hope. What is disturbing is the small number of scientists with large numbers of papers that were retracted on account of fraud. How were they able to get away with publishing so many papers, usually with fraudulent data, before the first identification of fraud that led to a checking of all their work? Ritchie (2020, 67–68))

cites, as an extreme example, the case of a Japanese anesthesiologist with 183 retracted papers.

More common are mistakes in data collection, unacknowledged sources of bias, hype, mistakes or biases in the handling of data and/or in data analysis, attaching a much higher degree of certainty to statistical evidence than the results warrant, and selection effects.

What gets published can be strongly affected by selection effects. There may be selection of a subset of data where there appears to be an effect of interest, choice of the outcome variable and/or analysis approach that most nearly gives the result that is wanted, and so on. In analysis of data from experiments where two treatments are compared, the common use of the arbitrary $p \leq 0.05$ criterion as a cutoff for deciding what will be published has the inevitable effect of selecting out for publication one in twenty (or more) results where there was no difference of consequence.[1]

9.2 The case of Eysenck and his collaborators

At the time of his death in 1997, Eysenck was the living psychologist most frequently cited in the peer-reviewed scientific literature. Much of his work was controversial in its time, with papers containing "questionable data and results so dramatic they beggared belief" (O'Grady 2020). He relied heavily on what has now been identified as heavily doctored data that was supplied to him by German collaborator Grossarth-Maticek. Particularly egregious was the claim that individuals with an identifiably cancer-prone personality had a risk of dying from cancer that was as much as 121 times higher than that of people with a "healthy" personality — one of several links that the duo claimed to have found between personality and mortality. Investigations into Eysenck's work, including collaborative work with Grossarth-Maticek, are ongoing. Fourteen papers have been retracted, and another 71 have received "expressions of concern". A large replication study conducted in 2004 found none of the claimed links, apart from a modest link between personality and cardiovascular disease.[2]

In a book published two years before his death, Eysenck made comments that provide an intriguing insight into his thinking.

[1] Note 2 (p. 89) in the notes at the end of the book makes relatively technical comments on common misunderstandings that affect the use of p-values.

[2] See further Craig, Pelosi, and Tourish (2021).

> Scientists have extremely high motivation to succeed in discovering the truth; their finest and most original discoveries are rejected by the vulgar mediocrities filling the ranks of orthodoxy. .. . . The figures don't quite fit, so why not fudge them a little bit to confound the infidels and unbelievers? Usually the genius is right, of course, and we may in retrospect excuse his childish games, but clearly this cannot be regarded as a licence for non-geniuses to foist their absurd beliefs on us. (Eysenck 1995, 197)

Eysenck was alluding to claims that Newton had manipulated data, and suggesting that it was excusable for other "geniuses" to do the same.

9.3 Detection of Covid-19 from chest images

Roberts et al. (2021) identified an astonishing 320 papers and preprints that appeared between 1 January 2020 to 3 October 2020, and which describe the use of new machine learning models for the diagnosis of COVID-19 from chest radiographic (CXR) and chest computed tomography (CT) images. Quality screening reduced this number to 62, which were then examined in more detail. None of the 62 satisfied these more detailed requirements, designed to check whether the algorithms used had been shown to be effective for use in clinical practice. Among other deficiencies, 48 did not complete any external validation, and 55 had a high risk of bias with respect to at least one of participants, predictors, outcomes and analysis. A significant number of systems were trained on X-rays from adults with covid-19 and children without, so that they were likely to detect whether the X-ray was from adult or child, rather than whether the person had Covid-19.

In an account of the results that appeared in New Scientist, Roberts (2021) comments that, relative to persisting to develop a model that will survive a rigorous checking process and might be used in practice, "it is far easier to develop a model with poor rigour and [apparent] excellent performance and publish this." This is a damning indictment of the way that large parts of the research and publication process currently work. The public good would be much better served by a process that encourages researchers to persist until it has been demonstrated that researchers have a model that meets standards such as are set out in Roberts (2021).

9.4 Laboratory studies — what do we find?

In the past several years, there has been a steady accumulation of evidence that relates to the claim, in Ioannidis (2005), that "most published research findings are false". Ioannidis has in mind, not published results in general, but primarily laboratory studies. Papers that have added to the body of evidence that broadly support claims made in the Ioannidis paper include:

- Amgen: Reproduced 6 only of 53 'landmark' cancer studies.
  - Begley and Ellis (2012)
  - Begley (2013) notes issues with the studies that failed
- Bayer: Main results from 19 of 65 'seminal' drug studies
  - NB, journal impact factor was not a good predictor!
  - Prinz, Schlange, and Asadullah (2011)
- fMRI studies: 57 of 134 papers (42%) had $\geq$ 1 case lacking check on separate test image. Another 14%, unclear ...
  - Kriegeskorte et al. (2009)

Issues that Begley (2013) notes are important both across the pre-clinical studies such as are his concern, and for clinical studies such as have been used, and are being used, the check the effectiveness and side effects of Covid-19 vaccines. Begley comments that the flaws many of the flaws that he identified "were identified and expunged from clinical studies decades ago". They include

- Were experiments blinded?
  - Was recording of all results done by another investigator who did not know what treatment had been applied?
- Were basic counts, measurements, and tests such as are done using Western blotting (used to detect specific proteins in a mixture) repeated?
- Were all results presented?
  - One concern is that when images are shown (e.g., of Western blot gels), images may be cropped in ways that more clearly identify the protein than is really the case.
- Were results shown for crucial control experiments?
  - Readers should be provided with evidence that there was unlikely to be any bias in the comparison of treatment with control.
- Were reagents validated?
  - Reagents must be shown to be fit for purpose, with results shown from analyses that validate their use.
- Were statistical tests appropriate?

  - Begley comments on common analysis flaws.

Begley comments

> What is also remarkable is that many of these flaws were identified and expunged from clinical studies decades ago. In such studies it is now the gold standard to blind investigators, include concurrent controls, rigorously apply statistical tests and analyse all patients — we cannot exclude patients because we do not like their outcomes.

Positive developments in the psychology community

The psychological science community is further advanced in addressing these issues that many other communities, with The Center for Open Science (COS) taking a strong lead in studies designed to document the extent of the issues.

Other Center for Open Science (COS) Projects have included

- The Reproducibility: Psychology project. This replicated 100 studies, from 100 journals in cognitive and social psychology.
  - In order to keep the graph simple, Figure 9.1 limits attention to the subset that relate to social psychology.
  - OSC (2015)
- Many Labs — reproduce 13 classical psych studies
  - Of 13 studies — 10: successful, 1: weakly, 2: no!
  - Plots show scatter across the 36 participating teams
  - Klein et al. (2014)
- Cancer Studies — 50 "most impactful" from 2010-2012
  - Kaiser (2015)
- Details from other relevant studies are given in the recent book Ritchie (2020) "Science fictions: Exposing fraud, bias, negligence and hype in science."

In the areas of science to which these studies relate, it is then clear that published claims that have not been replicated are as likely as not to be false. As Begley (2013) notes, however, there are clues that can allow a discerning to make a judgement on the credibility that should be given to claims that are made.

Figure 9.1 compares the effect size for the replicate with the effect size for the original, for the 54 social psychology studies included in the Reproducibility: Psychology project.

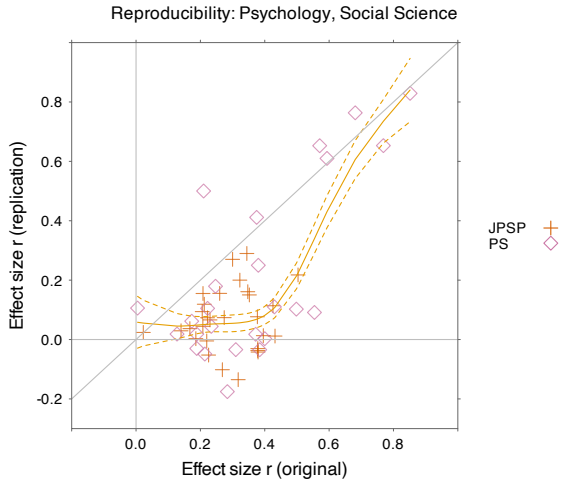

Figure 9.1: Psychology reproducibility project. Effect sizes are compared between the replication and the initial study, for the 54 social psychology studies included in the Reproducibility: Psychology project. The journals that are represented are Psychological Science (PS), and Journal of Personality and Social Psychology (JPSP)

The effect size is a measure of the average difference found, divided by an estimate of variability for the individual results. The effect size is smaller for the replication in 47 out of the 54 studies. A smooth curve, with confidence interval, has been fitted. It is only for an original effect size greater than 0.4 that one starts to see a positive correlation between the effect sizes for the replicate and that for the original.

Other replication studies

The critiques have limited relevance to areas where the nature of the work forces collaboration between scientists with diverse skills, widely across different research groups. Where all data and code used in modelling are out in the open for all to see and evaluate, and research requires co-operation internationally across multiple areas of expertise, ill-founded or exaggerated claims are unlikely to survive long. Differences in the extent to which the nature of the work force co-operation go a long way to explaining the large variation that is evident in

standards for published work.

Experimental results that have not been independently replicated should be treated as provisional, awaiting confirmation. Replication provides an indispensable check on all the processes involved. If there are gaps or inaccuracies in the report and/or supplementary material that make it impossible to replicate work, this soon becomes obvious. Another research group, especially one that has developed expertise in checking over the work of others, will not usually repeat the same mistakes.

The peer review process does at least impose some minimal checks on what is published. In some cases, issues may be identified subsequent to publication. The case is at least better than for claims made by those who promote "alternative medicines", nowadays often on the internet, offering "evidence" that is anecdotal.

9.5 Truths that special interests find inconvenient

The evidence that human induced greenhouse gas emissions are driving global warming has for the past two decades or more been overwhelming. Fierce criticism of any weakness in the evidence presented has, while delaying effective action, helped ensure that published work in this area meets unusually high standards.

Styles of argument

The tobacco industry has made extensive use of "the science is not settled" arguments in its efforts to dismiss the evidence that smoking causes lung cancer. The same PR firms, and the same researchers used to support these efforts were later used in the attempt to undermine climate science.[3] The goal has been to raise doubt, create confusion, and undermine the science.

In spite of the support that has been available from the fossil fuel industry for research that is critical of mainstream climate science research, no substantially different alternative account has emerged, and no climate change models have emerged that give results that differ widely from the consensus. Richard Muller's Berkeley Earth Surface Temperature Study (BEST) is interesting because Richard Muller had been known for his climate skepticism, in part based

[3] https://www.scientificamerican.com/article/tobacco-and-oil-industries-used-same-researchers-to-sway-public1/>

on legitimate scientific concerns. A large part of the funding came from the right-wing billionaire Charles Koch, known for funding climate skeptic groups such as the Heartland Institute.

Muller made headlines when he announced his acceptance of what climate scientists had been saying for more than 15 years previous:

> After years of denying global warming, physicist Richard Muller now says "global warming is real and humans are almost entirely the cause."[4]

The broad sweep of work in climate science is, because it has survived informed critique, and because of the diversity of contributing skills and data, unusually secure. Details, especially as they affect what may happen in individual countries, are subject to continual revision.

A particular issue is the role of the greenhouse gases, and of their interactions with water vapour. Their direct effect is greatly magnified by the consequent warming of the air in which water vapour is present, allowing it to retain more vapour and trap more heat). A standard denialist trump card has been to claim the authority of scientists who have standing in their own areas for the claim that the contribution of greenhouse gases is small relative to that of water vapour.[5] The scientists involved are among a number who have fossil fuel industry links, and have allowed themselves to be used as advocates for an industry that sees action on climate change as a threat.[6]

Big Pharma — inconvenient data

Between 1999 and 2019, opioid deaths in the United States increased by a factor of six, to almost 50,000. A large contributor has been the increased use of prescription opioids. Purdue Pharma stands out for its aggressive marketing of oxycodone, sold under the brand name OxyContin, arguing that concerns over addiction and other dangers from the drugs were overblown. In September 2019, Purdue Pharma declared bankruptcy, facing significant liability in OxyContin and opioid addiction lawsuits. Details of settlements are still being worked in

[4] https://courses.seas.harvard.edu/climate/eli/Courses/global-change-debates/Sources/Hockeystick-global-temperature/more/Richard-Muller/Muller-is-a-believer-Hallelujah.pdf>

[5] See https://west.web.unc.edu/climate-change/ and <>

[6] https://insideclimatenews.org/news/12032015/leaked-email-reveals-whos-who-list-climate-denialists-merchants-of-doubt-oreskes-fred-singer-marc-morano-steve-milloy

the courts.[7]

Strict regulations govern, in most countries, the approval of prescription drugs. Purdue Pharma exploited the much more limited control over off-label use of approved drugs, i.e., use for purposes for which they have not received formal approval, and for a time stayed under the radar.

Another scandal that demonstrates how drug companies can sometimes work their way around the approval process concerns the drug Vioxx, likewise marketed as a painkiller.(Valentine and Prakash 2007) Concerns that the drug might be increasing the risk of heart attacks began to emerge in the months following its approval for use in May 1999. By November 1, a study set up to investigate these concerns reported 79 heart attacks out of 4000 among those taking the drug, as opposed to 41 among a comparator group that was taking naproxen. The drug continued on the market as the evidence against Vioxx strengthened further. It was argued, with no evidence to back this up, that naproxen likely had a protective effect. In any case, why allow Vioxx to go to market when naproxen was clearly carried less risk.

In September 2004, when a colon-polyp prevention study showed that Vioxx increased the risk of heart attack after 18 months, Merck withdrew the drug. A Lancet paper that was published later estimated that between 88,000 and 140,000 Americans had heart attacks from taking Vioxx.[8] The increased risk continued long after patients had ceased taking the drug.

What can one trust?

One may hope that checks on any new drug will be stricter than was the case for Vioxx, that lessons have been learned, that where in future drugs come up for approval, checks will continue on possible long-term effects. With fringe and quack medicines, there are no such checks.

One can take encouragement from the way in which evidence of the type discussed, that one other drug is doing more harm than good, has in the cases discussed finally come to light. They illustrate the importance of collecting the relevant data, and of taking note of what the data have to say.

It has been interesting to follow approval processes for Covid-19 vaccines. At

[7] https://topclassactions.com/lawsuit-settlements/open-lawsuit-settlements/opioids/purdue-opioid-addiction-class-action-settlement/

[8] Jüni et al. (2004)

the time of writing, the Pfizer, Astrazena and Moderna vaccines have had, in addition to their testing in clinical trials, extraordinary levels of testing in clinical practice, with risks that are small relative to the small risk that we take ever time we cross a busy road. The trials, and the use of these vaccines in clinical practice, have been unusual in the levels of scrutiny that they have attracted from experts worldwide.

9.6 Tricks used to dismiss established results

The headings, and some of commentary, are adapted from an article by Associate Professor Hassan Vally from La Trobe University, that appeared in *The Conversation* for March 9 2021.

1. "The 'us versus them' narrative"
   The powers that be are trying to deceive us." Ask who is making the real attempt to deceive — commercial interest groups, peddlers of quack treatments, influence peddlers, . . .
2. 'I'm not a scientist, but…'
   Meaning perhaps: "I'm not a scientist, but that does not stop me making an authoritative pronouncement that flies in the face of established results." Or: "I know what the science says, but I'm keeping an open mind". Politicians are among the most frequent offenders.
3. Reference to 'the science not being settled'
   Much science is of course not settled, and one can then expect scientists to openly argue different points of view based on available evidence. Or, what has come to be accepted wisdom, may turn out to be wrong or in need of substantial revision. Challenges to the accepted wisdom have, however, to be carefully argued, and themselves survive informed critique.
4. Overly simplistic explanations
   Oversimplifications and generalizations are central to many anti-science arguments. Science is often messy, complex and full of nuance. The truth can be harder to explain, and sometimes sound less plausible, than a simple but incorrect explanation. Simplistic statistical arguments are common, e.g., fit a line to world average temperature data for the cherry-picked years 1998-2008, ignoring year to year correlation as well as influences that operate over longer time periods.
5. Cherry-picking
   One is not entitled to choose one study over another just because it aligns with what you prefer to believe. This is not how science works.

Not all studies are equal; some provide much stronger evidence than others. The way that choices are made has to stand up to critical scrutiny.

Chapter 10

Notes

1. The Jung et al. (2014) US hurricane data

Figures 8.7 and 8.8 checked for a difference between the fitted male and female line or curve. Jung et al's approach was, instead, to examine whether numbers of deaths varied with the "femaleness" of name, as judged by students in 2014.

As a check on how the popularity of a name for each of females and males may have changed with time, Table 10.1 uses US social security administration data to show numbers for names where the range of variation over 1950 to 2012 in the relative number of females was 0.08 or more.

In other cases, relative number of females were always either in the range 0 to 0.07 (i.e., mostly female), or in the range 0.93 to 1 (i.e., mostly female).

Figure 10.1 shows how the numbers of each sex changed over time, for the first

Table 10.1: The minimum and maximum value of the relative proportion of female names, and the difference, are shown for the eight names that showed the greatest change over the years 1950 to 2012.

	Fran	Charley	Cleo	Sandy	Erin	Inez	Carmen	Bret
Minimum	0	0.00	0.39	0.68	0.84	0.86	0.88	0.00
Maximum	1	0.87	0.94	0.95	0.99	1.00	0.98	0.08
Difference	1	0.87	0.55	0.28	0.15	0.14	0.10	0.08

six of the names in Table 10.1.

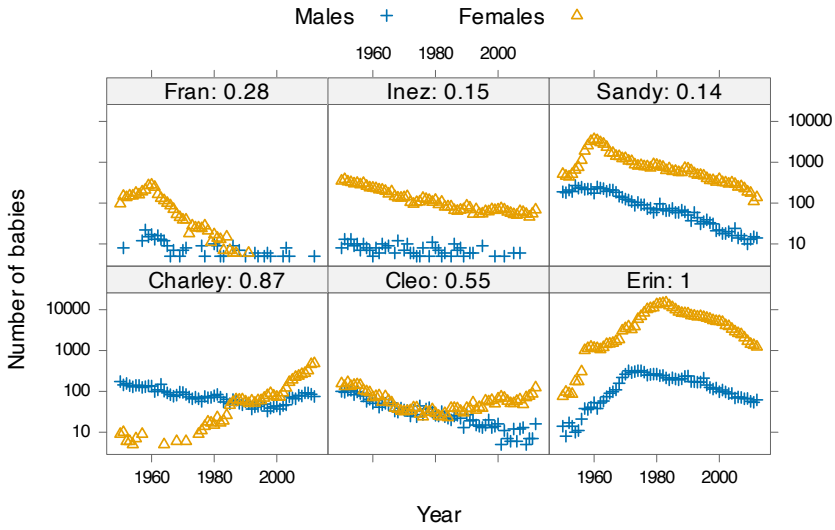

Figure 10.1: Change in numbers of names given to males and females over the years 1950 to 2012, for the six names where the maximum difference in relative frequency was more than 0.1. The maximum change is shown against each name.

The names where the choices of parents at the time are likely to most different from that for students in 2014 are those where there has been greatest change over time (i.e., especially, Charley, Cleo and Fran).

Other differences from the analyses on which Figures 8.7 and 8.8 were based are

- As the primary measure of the risk posed by the hurricanes, the authors used a 2013 US$ estimate of damage, intended for insurance purposes, for a comparable hurricane in 2013. Figure 8.7 uses what is surely the more relevant measure, namely `NDAM2014` — this converts the estimate of damage caused at the time to 2014 US$.[1]
- Jung et al allowed for minor effects from barometric pressure at landfall, and interactions. Again, these do not affect the conclusions reached.

[1]Fortuitously, this change makes no difference of consequence to the graph or to the conclusions reached.

2. *What does a p-value tell the experimenter?

P-values are widely used to indicate whether a difference. What follows introduces technicalities that have been avoided in the preceding chapters.

Consider an experiment that compares a treatment with a control. For example, does taking a drug of interest reduce sleeplessness? The starting point for a p-value calculation is the NULL hypothesis assumption of no difference in effect between treatment and control. A p-value measures the probability that a difference in measured effect as large as that observed, or larger would, assuming the NULL hypothesis, occur by chance. The definition says nothing about the p-value that can be expected if there is a difference. It does not give, as is sometimes thought, the probability that there really is no difference.

Issues for understanding the meaning of a p-value are:

- Finding $p \leq 0.05$ is all very well. What one really needs to know is that there is an alternative that is substantially more likely.
- $p = 0.05$, or close to 0.05, is at the upper end of the range of p-values that occur with a probability of 0.05. Treating $p = 0.05$ as an event that occurs with probability 0.05 under the NULL exaggerates the evidence that it provides against the NULL.
  - In fact, under the NULL, all values between 0 and 1 are equally likely!
- Prior probabilities matter. If for example the data is from one of a set of drug trials where 99 out of 100 can be expected to show no effect, the 99 out of 100 instances where there is a 1 in 20 change of $p <= 0.05$ have to be set again 1 out of 100 instances where there is a potential to detect an effect. A $p <= 0.05$ is more than $99/20 = 4.95$ times as likely to come from a drug with no effect as from one that has a real effect.
  - What size of effect is, based on whatever is known from past comparable investigations, likely? A reasonable default is to center the distribution on zero, with values that become increasingly less likely as the distance from zero increases. Under plausible assumptions of this general type, it can be shown that with p=0.05, the prior relative likelihood should be multiplied by, at most, around 2.5.[2]

[2]See https://statisticsbyjim.com/hypothesis-testing/interpreting-p-values/

Books, videos, and websites

Books referred to in the text

- Kahneman (2013) . Thinking, fast and slow.
  - Interview with Kahneman[3]
  - A brief animated overview of some key points[4]
- Smith (2014) . Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics
  - Brian Lehrer interview with Smith[5]
- Ellenberg (2015) . How not to be wrong.
  - There are links to several Ellenberg video clips[6]
- Levitin (2016) . A field guide to lies and statistics.
- Nisbett (2016) . Tools for smart thinking.
- Cairo (2013) . The functional art: an introduction to information graphics and visualization.
- Ritchie (2020) . Science fictions: Exposing fraud, bias, negligence and hype in science.

Videos

- Randomized controlled trials — animated summary of key issues[7]
- Yule-Simpson paradox — animated video[8]

[3] https://www.youtube.com/watch?v=PirFrDVRBo4
[4] https://www.youtube.com/watch?v=uqXVAo7dVRU&t=28s
[5] http://www.garysmithn.com/standard-deviations.html
[6] http://www.thelavinagency.com/news/new-videos-jordan-ellenberg-runs-the-numbers
[7] https://www.youtube.com/watch?v=Wy7qpJeozec
[8] https://www.youtube.com/watch?v=ZDinnCwP3dg

Websites

- BBC links to helpful web resources[9]
- The Cochrane Center[10]
- Harding Center for Risk Literacy[11]
- Winton Centre for Risk and Evidence Communication[12]

[9] http://www.bbc.co.uk/editorialguidelines/guidance/reporting-statistics
[10] https://www.cochrane.org/
[11] https://www.hardingcenter.de/en
[12] https://wintoncentre.maths.cam.ac.uk/

References

Anglemyer, Andrew, Hacsi T Horvath, and Lisa Bero. 2014. "Healthcare Outcomes Assessed with Observational Study Designs Compared with Those Assessed in Randomized Trials." *Cochrane Database of Systematic Reviews*, no. 4.

Arkes, Hal R, and Wolfgang Gaissmaier. 2012. "Psychological Research and the Prostate-Cancer Screening Controversy." *Psychological Science* 23 (6): 547–53.

Begley, C. G. 2013. "Reproducibility: Six Red Flags for Suspect Work." *Nature* 497 (7450): 433–34. https://doi.org/10.1038/497433a.

Begley, C. G., and L. M. Ellis. 2012. "Drug Development: Raise Standards for Preclinical Cancer Research." *Nature* 483 (7391): 531–33. https://doi.org/10.1038/483531a.

Boot, HM, and JH Maindonald. 2008. "New Estimates of Age-and Sex-Specific Earnings and the Male–Female Earnings Gap in the British Cotton Industry, 1833–1906 1." *The Economic History Review* 61 (2): 380–408.

Brawley, Otis W. 2018. "Prostate Cancer Screening: And the Pendulum Swings." Wiley Online Library.

Cairo, Alberto. 2013. *The Functional Art: An Introduction to Information Graphics and Vizualisation*. 1st ed. New Riders.

Chalmers, Iain, and Paul Glasziou. 2009. "Avoidable Waste in the Production and Reporting of Research Evidence." *The Lancet* 374 (9683): 86–89. https://doi.org/https://doi.org/10.1016/S0140-6736(09)60329-9.

Chisholm, Donna. 2016. "Birth Control." *The NZ Listener, October 8 - 14*, 18–24.

Cohen, P. 1996. "Pain Discriminates Between the Sexes." New Scientist, 2 November 1996.

Cokely, Edward T, Mirta Galesic, Eric Schulz, Saima Ghazal, and Rocio Garcia-

Retamero. 2012. "Measuring Risk Literacy: The Berlin Numeracy Test." *Judgment and Decision Making*.

Coleman, Thomas. 2019. "Causality in the Time of Cholera: John Snow as a Prototype for Causal Inference." *Available at SSRN 3262234*.

Collins, Jim. 2001. *Good to Great: Why Some Companies Make the Leap ... And Others Dont*. Random House.

Craig, Russell, Anthony Pelosi, and Dennis Tourish. 2021. "Research Misconduct Complaints and Institutional Logics: The Case of Hans Eysenck and the British Psychological Society." *Journal of Health Psychology* 26 (2): 296–311. https://doi.org/10.1177/1359105320963542.

Du Toit, George, Graham Roberts, Peter H Sayre, Henry T Bahnson, Suzana Radulovic, Alexandra F Santos, Helen A Brough, et al. 2015. "Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy." *N Engl J Med* 372: 803–13.

Ellenberg, Jordan. 2015. *How Not to Be Wrong*. 1st ed. Penguin Books.

Erik von Elm, MD, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, and Jan P Vandenbroucke. 2007. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *Ann Intern Med* 147 (8): 573e7.

Esserman, LJ, and WISDOM Study and Athena Investigators. 2017. "The WISDOM Study: Breaking the Deadlock in the Breast Cancer Screening Debate. NPJ Breast Cancer 3: 34." *NPJ Breast Cancer*. https://doi.org/10.1038/s41523-017-0035-5.

Eyler, John M. 1973. "William Farr on the Cholera: The Sanitarian's Disease Theory and the Statistician's Method." *Journal of the History of Medicine and Allied Sciences* 28 (2): 79–100.

———. 2004. "The Changing Assessments of John Snow's and William Farr's Cholera Studies." In *A History of Epidemiologic Methods and Concepts*, 129–39. Springer.

Eysenck, Hans Jürgen. 1995. *Genius: The Natural History of Creativity*. 12. Cambridge University Press.

Farquhar, Cynthia, Lesley McCowan, and S Fleming. 2016. "Letter to the Editor of PloSMedicine."

Fisher, Ronald Aylmer. 1937. *The Design of Experiments*. 2nd ed. Oliver; Boyd.

Fung, Jason. 2020. *The Cancer Code: A Revolutionary New Understanding of a Medical Mystery*. Harper Wave.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other*

Stories. Cambridge University Press.

Gøtzsche, Peter C, and Karsten Juhl Jørgensen. 2013. "Screening for Breast Cancer with Mammography." *Cochrane Database of Systematic Reviews*, no. 6.

Harris, Ian. 2016. *Surgery, the Ultimate Placebo: A Surgeon Cuts Through the Evidence*. NewSouth.

Hassall, Arthur Hill. 1850. "Memoir ON THE ORGANIC ANALYSIS OR MICROSCOPIC EXAMINATION OF WATER: Supplied to the Inhabitants of London and the Suburban Districts." *The Lancet* 55 (1382): 230–35.

Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." *Cambridge Journal of Economics*. Oxford University Press. http://www.peri.umass.edu/fileadmin/pdf/working_papers/working_papers_301-350/WP322.pdf.

Hu, Frank B. 2013. "Resolved: There Is Sufficient Scientific Evidence That Decreasing Sugar-Sweetened Beverage Consumption Will Reduce the Prevalence of Obesity and Obesity-Related Diseases." *Obesity Reviews* 14 (8): 606–19.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *CHANCE* 18 (4): 40–47. https://doi.org/10.1080/09332480.2005.10722754.

Jung, Kiju, Sharon Shavitt, Madhu Viswanathan, and Joseph M Hilbe. 2014. "Female Hurricanes Are Deadlier Than Male Hurricanes." *Proceedings of the National Academy of Sciences* 111 (24): 8782–87. http://www.pnas.org/cgi/doi/10.1073/pnas.1402786111.

Jüni, Peter, Linda Nartey, Stephan Reichenbach, Rebekka Sterchi, Paul A Dieppe, and Matthias Egger. 2004. "Risk of Cardiovascular Events and Rofecoxib: Cumulative Meta-Analysis." *The Lancet* 364 (9450): 2021–29.

Kahneman, Daniel. 2013. *Thinking, Fast and Slow*. 1st ed. Farrar, Straus; Giroux.

Kaiser, Jocelyn. 2015. "The Cancer Test." *Science* 348 (6242): 1411–13.

Klein, Richard A, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, et al. 2014. "Investigating Variation in Replicability." *Social Psychology* 45 (3): 142–52. https://doi.org/10.1027/1864-9335/a000178.

Kriegeskorte, Nikolaus, W Kyle Simmons, Patrick S F Bellgowan, and Chris I Baker. 2009. "Circular Analysis in Systems Neuroscience: The Dangers of Double Dipping." *Nature Neuroscience* 12 (5): 535–40. https://doi.org/10.1038/nn.2303.

Levitin, Daniel J. 2015. *The Organized Mind*. 1st ed. Penguin.

———. 2016. *A Field Guide to Lies and Statistics*. 1st ed. Penguin Random House.

Løberg, Magnus, Mette Lise Lousdal, Michael Bretthauer, and Mette Kalager. 2015. "Benefits and Harms of Mammography Screening." *Breast Cancer Research* 17 (1): 1–12.

Lord, Frederic M. 1967. "A Paradox in the Interpretation of Group Comparisons." *Psychological Bulletin* 68 (5): 304.

Martin, Richard M, Jenny L Donovan, Emma L Turner, Chris Metcalfe, Grace J Young, Eleanor I Walsh, J Athene Lane, et al. 2018. "Effect of a Low-Intensity PSA-Based Screening Intervention on Prostate Cancer Mortality: The CAP Randomized Clinical Trial." *JAMA* 319 (9): 883–95.

Moran, Patrick, and John Cullinan. 2022. "Is Mammography Screening an Effective Public Health Intervention? Evidence from a Natural Experiment." *Social Science & Medicine* 305: 115073. https://doi.org/https://doi.org/10.1016/j.socscimed.2022.115073.

Murphy, Caitlin C, Piera M Cirillo, Nickilou Y Krigbaum, Amit G Singal, Min-Jae Lee, Timothy Zaki, Ezra Burstein, and Barbara A Cohn. 2021. "Maternal Obesity, Pregnancy Weight Gain, and Birth Weight and Risk of Colorectal Cancer." *Gut*. https://doi.org/10.1136/gutjnl-2021-325001.

National Science Foundation, DC. National Science Board, Washington. 1975. *Science Indicators, 1974*. Superintendent of Documents.

Nisbett, Richard E. 2016. *Mindware. Tools for Smart Thinking*. 1st ed. Penguin Books.

O'Grady, Cathleen. 2020. "Famous Psychologist Faces Posthumous Reckoning." American Association for the Advancement of Science.

O'Neil, Cathy. 2016. *Weapons of Math Destruction*. 1st ed. Crown.

OSC. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716–16. https://doi.org/10.1126/science.aac4716.

Pashayan, Nora, Antonis C Antoniou, Urska Ivanus, Laura J Esserman, Douglas F Easton, David French, Gaby Sroczynski, et al. 2020. "Personalized Early Detection and Prevention of Breast Cancer: ENVISION Consensus Statement." *Nature Reviews Clinical Oncology* 17 (11): 687–705.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pearson, Karl, and Alice Lee. 1903. "On the Laws of Inheritance in Man: I. Inheritance of Physical Characters." *Biometrika* 2 (4): 357–462.

Popper, Karl Raimund. 1963. "Science: Problems, Aims, Responsibilities." *Federation of American Societies for Experimental Biology* 22: 961–72.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It

or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" *Nature Reviews Drug Discovery* 10 (9): 712–12. https://doi.org/10.1038/nrd3439-c1.

Raichand, Smriti, Adam G Dunn, Mei-Sing Ong, Florence T Bourgeois, Enrico Coiera, and Kenneth D Mandl. 2017. "Conclusions in Systematic Reviews of Mammography for Breast Cancer Screening and Associations with Review Design and Author Characteristics." *Systematic Reviews* 6 (1): 1–8. https://www.ncbi.nlm.nih.gov/pubmed/28532422.

Reinhart, Carmen M, and Kenneth S Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review* 100 (2): 573–78.

Ritchie, Stuart. 2020. *Science Fictions: Exposing Fraud, Bias, Negligence and Hype in Science*. Random House.

Roberts, Michael. 2021. "Artificial Intelligence Has Been of Little Use for Diagnosing Covid-19." New Scientist, 22 May 2021.

Roberts, Michael, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, et al. 2021. "Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans." *Nature Machine Intelligence*, no. 3335: 199–217.

Rosenbaum, P R. 2002. *Observational Studies*. 2nd ed. Springer.

Schulz, Kenneth F, Douglas G Altman, and David Moher. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ* 340. https://doi.org/10.1136/bmj.c332.

Secrist, H. 1933. *The Triumph of Mediocrity in Business*. Business Studies. Bureau of Business Research, Northwestern University. https://books.google.co.nz/books?id=n8RAAAAAIAAJ.

Seigworth, Gilbert R. 1980. "Bloodletting over the Centuries." *New York State Journal of Medicine* 80 (13): 2022–28.

Simon, Herbert A. 1992. "What Is an 'Explanation' of Behavior?" *Psychological Science* 3 (3): 150–61.

Smith, G. 2014. *Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics*. Duckworth Overlook.

Snow, John. 1855a. "On the Mode of Communication of Cholera." John Churchill.

———. 1855b. *On the Mode of Communication of Cholera*. John Churchill.

Soni, Payal D, Holly E Hartman, Robert T Dess, Ahmed Abugharib, Steven G Allen, Felix Y Feng, Anthony L Zietman, Reshma Jagsi, Matthew J Schipper, and Daniel E Spratt. 2019. "Comparison of Population-Based Observational Studies with Randomized Trials in Oncology." *Journal of Clinical Oncology*

37 (14): 1209.

Stark, Philip B, and Andrea Saltelli. 2018. "Cargo-Cult Statistics and Scientific Crisis." *Significance* 15 (4): 40–43.

Thaler, Richard H, and LJ Ganser. 2015. *Misbehaving: The Making of Behavioral Economics*. WW Norton New York.

Tu, Yu-Kang, and Mark S Gilthorpe. 2011. *Statistical Thinking in Epidemiology*. CRC Press.

Tversky, Amos, and Daniel Kahneman. 1983. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review* 90 (4): 293.

US Preventive Services Task Force. 1989. *Guide to Clinical Preventive Services: Report of the US Preventive Services Task Force*. Diane Publishing Company.

Valentine, Vikki, and Snigdha Prakash. 2007. "Timeline: The Rise and Fall of Vioxx." *NPR (National Public Radio)* 10.

Wainer, Howard. 2000. "Visual Revelations." *CHANCE* 13 (1): 47–48. https://doi.org/10.1080/09332480.2000.10542192.

Waterman, Robert H, and Thomas J Peters. 1982. *In Search of Excellence: Lessons from America's Best-Run Companies*. New York: Harper & Row.

Watkins, Stephen J. 2000. "Conviction by Mathematical Error?: Doctors and Lawyers Should Get Probability Theory Right." British Medical Journal Publishing Group.

Watson, J, A Adler, A Agweyu, et al. 2020. "Open Letter to MR Mehra, SS Desai, f Ruschitzka, and AN Patel, Authors of 'Hydroxychloroquine or Chloroquine with or Without a Macrolide for Treatment of COVID-19: A Multinational Registry Analysis'." *Lancet*, 31180–86.

Wernham, Ellie, Jason Gurney, James Stanley, Lis Ellison-Loschmann, and Diana Sarfati. 2016. "A Comparison of Midwife-Led and Medical-Led Models of Care and Their Relationship to Adverse Fetal and Neonatal Outcomes: A Retrospective Cohort Study in New Zealand." *PLOS Medicine* 13 (9): e1002134. https://doi.org/10.1371/journal.pmed.1002134.

Young, Scott WH. 2014. "Improving Library User Experience with a/b Testing: Principles and Process." *Weave: Journal of Library User Experience* 1 (1).

About the author

For the major part of his career, John Maindonald has worked with other researchers as a quantitative problem solver. He has held positions at Victoria University of Wellington, in DSIR, in HortResearch Crown Research Institute (now Plant and Food), and at The Australian National University (ANU). He joined the then newly formed ANU Centre for Bioinformation Science in 2001, formally retiring in 2005. Between 1983 and 1996, and occasionally after 1996, he reviewed the statistical content of numerous papers that appeared in DSIR (later, Royal Society) journals, notably the New Zealand Journal of Agricultural Research and the New Zealand Journal of Crop and Horticultural Research.

Between 2003 and 2015, he fronted 35 short courses (one week, or less) across different parts of Australia that demonstrated the use of the open source R data analysis and graphics system for data analysis purposes. He is the author of a book on Statistical Computation, and the senior author of the texts

- *Data Analysis and Graphics Using R — An Example-Based Approach*, Maindonald and Braun, Cambridge University Press 2010. This has sold more than 11,000 copies over the three editions.

- *A Practical Guide to Data Analysis Using R – An Example Based Approach*, Maindonald, Braun, Andrews. Cambridge University Press 2024 (in press).